Medical University of South Carolina

MEDICA

MUSC Theses and Dissertations

2019

Unified Approaches for Frequentist and Bayesian Methods in Two-Sample Clinical Trials with Binary Endpoints

Zhenning Yu
Medical University of South Carolina

Follow this and additional works at: https://medica-musc.researchcommons.org/theses

Recommended Citation

Yu, Zhenning, "Unified Approaches for Frequentist and Bayesian Methods in Two-Sample Clinical Trials with Binary Endpoints" (2019). *MUSC Theses and Dissertations*. 262.
https://medica-musc.researchcommons.org/theses/262

This Dissertation is brought to you for free and open access by MEDICA. It has been accepted for inclusion in MUSC Theses and Dissertations by an authorized administrator of MEDICA. For more information, please contact medica@musc.edu.

Unified Approaches for Frequentist and Bayesian Methods in Two-Sample Clinical Trials with Binary Endpoints

By
Zhenning Yu

A dissertation submitted to the faulty of the Medical University of South Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the College of Graduate Studies.

Department of Public Health Sciences
2019.07

Approved By:

_____
Caitlyn Meinzer, Ph.D.
Co-Chair, Advisory Committee

_____
Viswanathan Ramakrishnan, Ph.D.
Co-Chair, Advisory Committee

_____
Renee' Hebert Martin, Ph.D.

_____
Brian Neelon, Ph.D.

_____
Byron Gajewski, Ph.D.

_____
Lee Schwamm, M.D.

TABLE OF CONTENTS

DISSERTATION COMMITTEE MEMBERS

Caitlyn Meinzer, Ph.D.
Assistant Professor of Biostatistics
Department of Public Health Sciences,
Medical University of South Carolina

Viswanathan Ramakrishnan, Ph.D.
Professor of Biostatistics
Department of Public Health Sciences,
Medical University of South Carolina

Renee' Hebert Martin, Ph.D.
Associate Professor of Biostatistics
Department of Public Health Sciences,
Medical University of South Carolina

Brian Neelon, Ph.D.
Associate Professor of Biostatistics
Department of Public Health Sciences,
Medical University of South Carolina

Byron Gajewski, Ph.D.
Professor of Biostatistics
Department of Biostatistics,
The University of Kansas Cancer Center

Lee Schwamm, M.D.
Vice-Chairman
Department of Neurology,
Massachusetts General Hospital

ACKNOWLEDGMENTS

First and foremost, I would like to thank my two mentors, Dr. Caitlyn Meinzer, and Dr. Viswanathan Ramakrishnan, who supported me and guided me through this wonderful doctoral journal. As I began at the MUSC as an international student, Dr. Meinzer and Dr. Ramesh had not only provided valuable guidance for me to grow professionally as an academic researcher, but also helped me find my feet in Charleston as a foreigner. I would also thank my committee members, Dr. Martin, Dr. Neelon, Dr. Gajewski and Dr. Schwamm for their selfless efforts to improve my dissertation work. I want to acknowledge Dr. Wenle Zhao and Dr. Yuko Palesch for funding and opportunities to work closely with clinical trials, and Dr. Dongjun Chung and Dr. Gary Hardiman for providing me chances to work on other interesting research projects.

Of course, I could not make these achievements without my family and friends. I would like to thank my parents who had been encouraging me during the entire journal. I felt very fortunate to have such a great family. Additionally, I would love to express my greatest gratitude for Cass, whose optimism and positive mind often inspired me to light a candle when facing difficulties. I would also like to thank other family members and sincere friends who were my cheerleaders all the time. I love you guys so much!

Lastly, I would thank the Data Coordination Unit, the Department of Public Health Sciences at the MUSC. The collaborative environment shaped my future and developed my critical thinking skills. I also want to say special thanks to Dr. Michael Madson from the Writing Center who had been extremely patient to me and helping me consistently on improving this dissertation.

ABSTRACT

Two opposing paradigms, analyses via frequentist or Bayesian methods, dominate the statistical literature. Most commonly, frequentist approaches have been used to design and analyze clinical trials, though Bayesian techniques are becoming increasingly popular. However, these two paradigms can generate divergent results even in analyses of the same trial data, which may harm the scientific interpretability of the trial. Therefore, it is crucial to harmonize analyses under each approach. In this dissertation, novel unified approaches for one-sided frequentist and Bayesian hypothesis testing problems comparing two proportions in fixed-sample and group-sequential clinical trials are proposed. When a frequentist design with desired type I and II error rates are given, the unification is achieved by deriving specific Bayesian decision thresholds and sample sizes. Similarly, when a Bayesian design is given, the unification is achieved by deriving corresponding frequentist characteristics. In addition, theoretical methods to determine the Bayesian decision threshold, sample size and power are provided. Numerical results show that the unified approach can yield the same type I and II error rates for frequentist and Bayesian hypothesis tests through a numerical study. Further, detailed evaluations suggest that Bayesian priors specifications, allocation ratios, number of analyses can affect the resulting Bayesian sample sizes and decision thresholds. Overall, the unified approach can be adopted into the current clinical trial setting and is helpful to make trial results translatable between frequentist and Bayesian methods.

Keywords*: clinical trials, Bayesian design, frequentist design, hypothesis testing, unification, sample size calculation*

CHAPTER ONE: INTRODUCTION

1.1 Introduction

Frequentist and Bayesian approaches are two opposing paradigms commonly used in statistical inferences. The root of the disagreement between these two approaches lies in their fundamental philosophies. For example, in a clinical trial, clinical trialists are interested in finding the probability of success for a new treatment (i.e, the treatment effect). Frequentist methods, or called the classical methods, consider the treatment on each patient as an independent experiment and repeat the experiment for multiple times by enrolling a number of patients. The average success rate of multiple experiments approximates the true value of the probability of success for the new treatment.

In contrast, Bayesian approaches consider the probability of success as an uncertain quantity and associate the treatment effect with a probability distribution at the beginning of the trial. Those probability distributions are called prior distributions, as they often represent a priori knowledge about the effect the new treatment would have. When clinical trial data is collected, Bayesian methods update the probability distribution for treatment effect with evidence observed from the data. The evidence is also called the data likelihood, which represents the amount of information provided by data on the treatment effect. The updated probability distribution is referred to as the posterior distribution.

The fundamental disagreement between frequentist and Bayesian paradigms has led to a philosophical debate on which paradigm is superior for decades. Although it is mentioned in many publications, that frequentist approaches have dominated the statistical analysis since the early 20th century (Feigelson, Loredo, & Building, 1992; Greenland, 2006; Gupta, 2012; Kass, 2011; J. J. Lee & Chu, 2012). The early popularity of frequentist methods is largely due to frequentist

probability calculations do not involve many computations, whereas Bayesian methods can have complicated calculations and often demand computer powers that are unattainable in early days (J. J. Lee & Chu, 2012). To be more specific, frequentist probability calculations usually have closed-form solutions that allow quickly being computed, whereas Bayesian probability calculations only have closed-form solution under certain assumptions (e.g. the conjugate priors, for which the prior and the posterior distribution follows the same family of distribution). In addition, Bayesian methods can have hierarchical structures, which introduce uncertainty to parameters of the prior distribution. As a result, Bayesian probability calculations can be computationally intensive.

In clinical trials, frequentist methods have been the dominant paradigm since the first modern trial carried out in the 1940s (Gupta, 2012; J. J. Lee & Chu, 2012). Over the past decades, the medical community and regulatory agencies have adopted frequentist approaches as the testing standard in clinical trials (Teira, 2011). However, recent computational advance in software and computational power (e.g. Markov-Chain Monte-Carlo and WinBUGS) greatly facilitate the use of Bayesian techniques in statistical design and analysis including clinical trials. Formal acceptance from by the regulatory agencies (FDA, 2014; ICH, 2017) and well established Bayesian standard methods (S. M. Berry, Carlin, Lee, & Muller, 2010) also help popularize Bayesian methods in clinical trials.

There are several purported advantages of using Bayesian approaches in clinical trials. First, Bayesian results are more related to clinician's question than frequentist results. For example, a clinician may ask, 'How likely is that our experimental treatment to be better than the control?', or 'What is the probability of our null hypothesis to be true?'. Frequentist statisticians draw conclusions for the hypothesis test based upon the significance test and the *p*-value. However, *p*-values are not intuitive and are often misinterpreted as the probability of the null hypothesis to be

true by clinicians (Cohen, 2011; Goodman, 2008). In fact, the *p*-value should be interpreted as the probability of obtaining a treatment effect at least as extreme as the observed effect when the null hypothesis is true. Therefore, the *p*-value essentially is not answering the clinician's questions.

In contrast, Bayesian methods allow to directly assign a probability distribution to the null hypothesis or a statement such as the experimental arm is superior to the control arm(D. A. Berry, 2006). Thus, the definition of Bayesian probability is more intuitive in comparison with the frequentist definition of probability. Furthermore, The subsequent Bayesian updating and decision making is also an inherently sequential process (Bayarri & Berger, 2004). In addition, Bayesian methods enable to incorporate prior information for assessing treatment effects in clinical trials. Historical data from previous trials in a similar setting can provide valuable information and possibly reduce the sample size for clinical trials. For instance, Laptook et al. (2017) used a Bayesian analysis of the treatment effect of hypothermia in infants with hypoxic-ischemic encephalopathy (HIE). Because of the limited number of infants expected to enroll in their clinical trial, it is difficult to conduct a traditional statistical significance test. To resolve this problem, the authors incorporated historical data to yield a quantitative summary of the trial results.

Nevertheless, there are controversial opinions about the use of Bayesian priors. On the other hand, Fayers et al. (1997) had made justifications for the use of prior in Bayesian approaches. The authors suggested that a pessimistic prior can reduce possibilities to wrongly draw a conclusion that there is overwhelm treatment effect. As such, an optimistic prior can lower the chances that an effective treatment is early terminated for futility. On the other hand, some clinical trialists are concerned with choosing appropriate prior distributions as selecting a prior can be subjective and biased (J. J. Lee & Chu, 2012). This is because clinical trialists often determine the

prior distribution based on their own knowledge about a treatment effect. As a result, people may end up using different priors on the same dataset and draw different conclusions.

Discordance between conclusions makes trial results hard to interpret. Irony (2008) also found that the Bayesian inference was very sensitive to the prior variance addressing the cross-variability of historical studies. If there is only one historical study, clinical trialists cannot precisely estimate the variance. Consequently, the prior information may lead to wrong posterior inference and decisions. Regulatory agencies, therefore, preclude the use of informative priors when there is not enough historical data (Teira, 2011).

Gönen (2009) and Grieve (2016) have also discussed problems with available software in addition to prior selections. Lack of accessible user-friendly software can often discourage biostatisticians to use Bayesian methods because it is impossible for biostatisticians to write complex code every time to implement a Bayesian clinical trial design. Moreover, regulatory authorities including the FDA often require frequentist concepts, such as type I and II error rates, to be rigorously controlled (D. A. Berry, 2006; FDA, 2014). However, the type I and II error rates are frequentist concepts and are not naturally considered Bayesian philosophies. This is because that type I and II error rates measure probabilities of making wrong conclusions about the hypotheses while Bayesian methods directly measure the probability of the hypothesis. Therefore, many Bayesian designs are carried out without explicit consideration of frequentist characteristics. Very often, clinical trialists have to perform simulations to adjust Bayesian tuning parameters to satisfy the respective type I and II error rates. However, tuning those parameters can influence the optimality of original Bayesian designs (Ventz & Trippa, 2014).

Another problem is frequentist and Bayesian methods may differ in conclusions. When informative priors are used in Bayesian designs, clinical trialists may reach divergent conclusions

between Bayesian and frequentist inferences, even on the same data (FDA, 2014). The divergence harms the scientific merit of a clinical trial, as conclusions should be consistent across analytic methods. Consequently, there has been a long-time debate between frequentist and Bayesian clinical trialists about which one is advantageous to the other (Efron, 2005; D J Spiegelhalter, Freedman, & Parmar, 1994).

Thus, it is important to consider unifying Bayesian and frequentist methods in clinical trials. Notably, statisticians have put a lot of effort to seek parallels between frequentist and Bayesian methods. Several statisticians have reconciled or calibrated one of the methods to the other for data analysis to reach the same conclusion. For instance, a prominent contribution is that Casella and Berger (1989) showed adjusting priors in Bayesian approaches can lead to the same inference as frequentist methods when analyzing data. However, such prior calibrations during the data analysis are inappropriate for clinical trials. In clinical trials, the statistical analysis is well planned at the design stage so it must be analyzed as planned during analysis. As such, many other methods of unifying frequentist and Bayesian methods are not applicable to clinical trials because of adjusting parameters during analysis. Therefore, it is crucial to consider the problem from the design stage to unify frequentist and Bayesian methods in clinical trials.

Several biostatisticians have made contributions regarding harmonizing frequentist and Bayesian methods at the design stage. Inoue, Berry, and Parmigiani (2005) considered unifying frequentist and Bayesian methods at the design stage. The authors found a correspondence between Bayesian and frequentist methods to generate the same sample size. However, the authors defined a specific classification error for Bayesian hypothesis testing rather than using type I and II error rates. The classification error makes it difficult to translate results between the frequentist and Bayesian approaches.

More recently, Zhu et al. (2015) considered a simulation-based approach to introduce alpha spending functions (DeMets & Lan, 1994; Lan & Demets, 1983) into Bayesian group sequential trials. The authors showed that the overall type I error rate is controlled under multiple types of alpha spending stopping boundaries. However, Zhu et al. required sample size to be pre-specified for calculating stopping boundaries, which conflicted with most trial planning processes where the sample size is determined once stopping boundaries are derived. Furthermore, the authors did not consider conditional distributions of posterior probabilities of multiple analyses, nor varying futility boundaries, which were often found in realistic group-sequential trials.

Shi and Yin (2019) proposed another method to control type I error rate in single boundary Bayesian group sequential trials. The method reduced the computation burden required by simulations and maintained the desired type I error rate. However, the authors did not consider stopping the trial for futility at the interim analysis. Although both Zhu et al. and Shi and Yin reconciled frequentist methods to Bayesian methods to some extent, the final results were sensitive to Bayesian prior specifications. To be more specific, the Bayesian type I error rate can vary because stopping boundaries are computed without taking prior distributions into account. Furthermore, the other frequentist concept, the type II error rate, may not be preserved as desired either, under different prior distributions. Overall, there is no available method to unify frequentist and Bayesian approaches on both type I and II error rates in clinical trials.

To address the above issues and create an exact one-to-one mapping between frequentist and Bayesian group sequential methods given different Bayesian priors, we propose novel unified approaches for frequentist and Bayesian hypothesis testing problems in one-sided two-arm fixed-sample and group-sequential clinical trials with binary endpoints. When a frequentist design is given, the unified approach determines the Bayesian sample size and decision thresholds through

a theoretical approach. When a Bayesian design is given, the unified approach calculates the frequentist type I and II error rates, leading to a corresponding frequentist design. For group-sequential trials, alpha spending functions for controlling the overall type I error rate and beta spending function for controlling the overall type II error rate are utilized. It is assumed that clinical trials have two arms with binary outcomes. Beta conjugate priors and decision makings based on posterior probabilities for treatment difference are used in Bayesian methods. Also, we derive the distributions of Bayesian posterior probabilities and provide closed-form solutions to compute Bayesian type I and II error rates in this dissertation work. Additionally, a user-friendly software application implementing the proposed unified approaches has been built to allow ease of use by clinical trialists.

Using the proposed unified approaches, a detailed numerical investigation is performed to compare frequentist and Bayesian approaches for different design parameters, such as Bayesian prior specifications, numbers of analyses, allocation ratios and stopping boundaries. The comparison is objective because the frequentist methods and Bayesian methods correspond to each other regarding the type I and II error rates. The influences of these design parameters on the Bayesian maximum and expected sample sizes are also evaluated. Evaluation results can help clinical trialists to gain insights into the optimality of frequentist and Bayesian paradigms in clinical trials.

Overall, the dissertation work has the following highlights:

1. The dissertation establishes a 1-to-1 mapping between frequentist and Bayesian methods in clinical trials. This mapping achieves the unification of frequentist and Bayesian without adjusting any tuning parameters. Therefore, the unified

approaches can lower the entry point to Bayesian methods for some frequentist clinical trialists, as well as help Bayesian biostatisticians embrace classical methods.

2. The unified approaches inherit the classical frequentist approach for fixed sample clinical trials and group-sequential clinical trials(Jennison & Turnbull, 2000; Lachin, 1981), which is conceptually easy to frequentist clinical trialists. From the perspectives from Bayesian biostatisticians, The proposed theoretical methods can help to rid of long-time simulations and obtain frequentist characteristics quickly.
3. Further evaluations show that frequentist and Bayesian methods can outperform each other with respect to sample sizes under the circumstance. Results indicate that frequentist and Bayesian methods both can play important roles in clinical trials. Thus, the dissertation can help clinical trialists to select the optimal method regarding the sample sizes in real practice.

Note, before developing novel unified approaches for frequentist and Bayesian methods in clinical trials, a careful literature review is conducted to help crystalize our specific aims. The entire process of the literature review is displayed in **Figure 1.1** in appendix section **1.5**. The review starts with understanding frequentist and Bayesian methods and their conflicts. Then the advantages and disadvantages of the Bayesian approach are reviewed. Finally, existing approaches to reconciling or unifying frequentist and Bayesian methods are reviewed.

The rest of this chapter summarizes the literature review results and is structured as follows. **Chapter 1.2** and **1.3**, briefly introduce frequentist and Bayesian methods, with some advantages and pitfalls which have been stated in the literature. Most of the frequentist and Bayesian methods introduced are used to develop the novel unified approaches in the following chapters. In **Chapter**

1.4, a detailed review summarizes previous work on exploring correspondences between the frequentist and Bayesian methods is present.

1.2 Frequentist Methods

1.2.1 Frequentist Philosophy

The frequentist paradigm originates from the concepts of *p*-value and hypothesis testing. The *p*-value was proposed by Ronald Fisher in the 1920s as indices of the strength of the evidence against the null hypothesis; it represents the probability of obtaining data equal to or more extreme than the observed one if the null is true. Commonly, statisticians degrade the *p*-value into dichotomy based on some decision threshold. Fisher suggested a *p*-value of 0.05 or less as indicating strong evidence against the null.

Instead, Neyman and Pearson proposed a hypothesis-testing framework, in which they defined type I and II error rates, denoted as α and β, and a decision threshold (or called the critical value by frequentist statisticians), denoted as *c*. The type I error rate is defined as the probability to accept an alternative hypothesis when the null hypothesis is true, whereas the type II error rate is defined as the probability of failing to reject the null when the alternative is true. For the given type I and II error rates, the decision threshold can be specified. In the hypothesis testing framework, a test statistic, denoted as T, is calculated in each hypothesis test. The null hypothesis is rejected if the test statistic T is greater than the pre-specified decision threshold *c*. If the test statistic falls into above the decision threshold, the null hypothesis is rejected. Although *p* value and Neyman-Pearson hypothesis testing are different concepts, current frequentist methods combine them together in practice. For instance, one can reject the null in a normal test at a significance level of 0.05, either using the decision threshold of ± 1.96 or the *p* value of 0.05.

1.2.2 Frequentist Hypothesis Testing

Typically, a frequentist hypothesis test involves a null hypothesis, denoted as H_0, and an alternative hypothesis, denoted as H_1. In clinical trials comparing success rates (denoted as π_1 and π_2) between two treatments, the H_0 is often stated such that success rates of the experimental and control groups are not different (i.e. $H_0: \pi_1 = \pi_2$). The alternative hypothesis can be one-sided or two sided. In one-sided tests, the H_1 is stated such that the success rate of the experimental arm is greater than that of the control (i.e. $H_1: \pi_1 > \pi_2$) or the mortality rate of the experimental arm is smaller than that of the control (i.e. $H_1: \pi_1 < \pi_2$). In two-sided tests, the H_1 is stated such that the success rate of the experimental arm is either greater or smaller than that of the control (i.e. $H_1: \pi_1 \neq \pi_2$). Sometimes, the null hypothesis can be stated also as the rate difference between the experimental arm and the control arm is less than a minimal clinical meaningful significance, Δ. That is, $H_0: \pi_1 - \pi_2 \leq \Delta$. Correspondingly, the alternative hypothesis can be stated as success rates between the experimental arm and the control arm is greater than Δ, $H_1: \pi_1 - \pi_2 > \Delta$.

When performing a hypothesis test, frequentist philosophy considers probabilities as frequencies. That is, π_1 and π_2, are meaningful only when they are considered as the limit of a specific event's (e.g. success, mortality) relative frequency in repeated trial experiments:

$$\pi_i = \lim_{n_i \to \infty} \frac{y_i}{n_i}, i = 1,2,$$

where the subscript i is used to denote treatment groups, y_i and n_i are the number of events and number of trials. It can be seen that π_1 and π_2 are fixed numbers and have no associated probability distributions. Since it is impossible to conduct endless experiments and obtain the value of π, the mean of the sample, denoted as $\hat{\pi}$, is often used to approximate π. $\hat{\pi}$, however, has a distribution conditional on the value of π.

Furthermore, assume there are multiple sample means, $\hat{\pi}_{.1}$, $\hat{\pi}_{.2}, \ldots, \hat{\pi}_{.m}$ of m samples collected from the population, the frequentist central limit theorem states that the sampling distribution follows a normal distribution regardless of what the original distribution π has. The resulting normal distribution makes it convenient to perform the frequentist hypothesis test and sample size calculation. For the hypothesis testing problem, the test statistic T is constructed based on the sampling distribution. For instance, assume there is a clinical trial comparing two treatments, in which patients are equally randomized into two groups. The test statistic is defined as:

$$T = \frac{\widehat{\pi_1} - \widehat{\pi_2}}{\sqrt{\frac{\widehat{\pi_1}(1-\widehat{\pi_1})}{n_1} + \frac{\widehat{\pi_2}(1-\widehat{\pi_2})}{n_2}}}.$$

Using the Neyman-Pearson hypothesis testing framework, the decision threshold *c* is defined as $Z_{1-\alpha}$, where $Z_{1-\alpha}$ is the $(1-\alpha)th$ quantile of the standard normal distribution. Alternatively, we can calculate the p value by subtracting the percentile of T evaluated at the standard normal distribution from one, and compare the *p* value to a threshold value, e.g. 0.05. The threshold is also called the significance level.

1.2.3 Freuqentist Sample Size Determination for Fixed-Sample Methods

Similar to the hypothesis testing problem, the frequentist sample size determination is based on the sampling distribution. Methods for calculating sample size for fixed-sample clinical trials comparing two proportions was illustrated in detail by Lachin (1981). The frequentist sample size calculation approach utilizes the relationship between type I and II error rates and the decision threshold (see **Figure 1.2**). Mostly, the sample size is chosen as the smallest value to control some pre-specified type I and II error rates. Consider the same clinical trial comparing two equal-sized treatment groups as in **section 1.2.2**, and let P_1 and P_2 denote the assumed treatment effect for two groups at the beginning of the trial. The sample size for each treatment group is calculated as:

$$n_i = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{P_1 - P_1}\right)^2 \left(P_1(1-P_1) + P_2(1-P_2)\right), i = 1,2,$$

where $Z_{1-\alpha}$ and $Z_{1-\beta}$ are the $(1-\alpha)th$ and $(1-\beta)th$ quantile of the standard normal distribution. If patients are unequally assigned to two treatment arms with an allocation ratio $r, r \neq 1$, so that the patient size in the experimental arm is r times the size in the control arm. The sample sizes for the two treatment can be calculated as:

$$\begin{cases} n_1 = rn_2 \\ n_2 = \left(\frac{Z_{1-\alpha} + Z_{1-\beta}}{P_1 - P_1}\right)^2 \left(\frac{P_1(1-P_1)}{r} + P_2(1-P_2)\right) \end{cases}.$$

Overall, the hypothesis testing and sample size determination are conceptually simple and the central limit theorem allows quick computation for results. Thus, frequentist approaches have been the predominant methods for modern clinical trial research since the 1940s (J. J. Lee & Chu, 2012). The medical community and regulatory agencies have also adopted frequentist approaches as the testing standard in clinical trials (Teira, 2011). In addition, the ICH E-9 guidance largely refers to the use of frequentist methods when discussing hypothesis testing problems (ICH, 2017).

1.2.3 Group-Sequential Methods

For the above frequentist hypothesis tests and sample size determination for clinical trials, one assumption is that those clinical trials have only a final analysis when all patients recruited. Such fixed-sample design is unjustified for ethical and financial concerns because patients entered the study sequentially and the trial should be terminated when accumulated data is sufficient to conclude (Pocock, 1977). Thus, frequentist group sequential methods are proposed to analyze clinical trial data (Peter Armitage, 1958; O'Brien & Fleming, 1979; Pocock, 1977). Armitage (1958) was the first person to introduce sequential method into clinical trials, and Pocock (1977)

and O'Brien and Fleming (1979) help polish the frequentist group-sequential methods for clinical trials. In group sequential clinical trials, the patients' enrollment is partitioned into multiple stages. At the end of each stage, a hypothesis test is carried out on accumulating clinical trial data, where all statistical tests performed prior to the complete patient enrollment are called interim analyses. Suppose there are J stages and let j denote the stage, $j = 1, \dots, J$. Let $n_{.j}$ denote the accumulated number of patients at the end of the jth stage. A hypothesis test at the jth stage is formalized as follow:

$$T_j = \frac{\widehat{\pi_{1J}} - \widehat{\pi_{2J}}}{\sqrt{\frac{\widehat{\pi_{1J}}(1-\widehat{\pi_{1J}})}{n_{1j}} + \frac{\widehat{\pi_{2J}}(1-\widehat{\pi_{J}})}{n_{2j}}}}.$$

Note, $T_1, T_2, \dots, T_J$ follow a multivariate normal distribution (Jennison & Turnbull, 2000). That is, $T_j \sim N\left((P_1 - P_2)\sqrt{I_j}, 1\right)$, where $(P_1 - P_2)\sqrt{I_j}$ is called a drift parameter and I_j is the Fisher information available at the jth stage. Further, the covariance between T_{j_1} and T_{j_2}, $j_1 < j_2$ is $\sqrt{I_{j_1}/I_{j_2}}$.

In addition, the group-sequential design can have two decision thresholds for each analysis, one for efficacy, denoted as $c_{E,.}$ and another for futility, denoted as $c_{F,.}$. The set of $c_{E,j}$ and $c_{F,j}$, $j = 1, \dots, J$, are also called efficacy and futility stopping boundaries. The trial is stopped at the jth stage for overwhelming benefits of the experimental treatment if $T_j > c_{E,j}$, or it is stopped for sufficient negative treatment effect if $T_j < c_{F,j}$. Otherwise, the trial shall continue to the next analysis. Note, a final conclusion must be made at the final analysis, so that $c_{E,J}$ is set to equal $c_{F,J}$. Such group-sequential methods are ethically and economically beneficial, especially for confirmatory trials in which patient sample sizes are large.

Furthermore, these stopping boundaries are computed to achieve the desired type I and II error rates. Since repeated analyses are conducted in group-sequential clinical trials, it is widely acknowledged that the allowed type I error rate for each analysis has to be smaller than the desired type I error rate in order to preserve the overall type I error rate (Armitage, McPherson, & Rowe, 1969; DeMets & Lan, 1994; O'Brien & Fleming, 1979; Pocock, 1977). The overall type I error rate is usually the sum of the stagewise type I error rates and the same for the overall type II error rate. For instance, in group sequential trials with only efficacy boundaries, the following relationship between a set of efficacy boundary values and the overall type I error rate should be met:

$$\sum_{j=1}^{J} Pr_{H_0}\left(\bigcap_{j^*}^{j-1} T_{j^*} \leq c_{E,j^*} \cap T_j > c_{E,j}\right) = \alpha,$$

and for group sequential trials with only futility boundaries, the following relationship should be met:

$$\sum_{j=1}^{J} Pr_{H_1}\left(\bigcap_{j^*}^{j-1} T_{j^*} \geq c_{F,j^*} \cap T_j < c_{F,j}\right) = \beta.$$

For group sequential clinical trials with both efficacy and futility boundaries, the following equations should be satisfied for stopping boundaries and type I and II error rates:

$$\sum_{j=1}^{J} Pr_{H_0}\left(\bigcap_{j^*}^{j-1} c_{F,j^*} \leq T_{j^*} \leq c_{E,j^*} \cap T_j > c_{E,j}\right) = \alpha,$$

and

$$\sum_{j=1}^{J} Pr_{H_1}\left(\bigcap_{j^*}^{j-1} c_{F,j^*} \leq T_{j^*} \leq c_{E,j^*} \cap T_j \leq c_{F,j}\right) = \beta.$$

Note that all the above conditions assume binding boundaries where the trial is stopped if futility boundaries have been crossed during interim analyses. Some clinical trialists prefer using non-binding boundaries where flexibility is offered to interim analyses (Schüler, Kieser, & Rauch, 2017). The trial can continue despite that futility boundaries are crossed and efficacy boundaries are derived regardless of stopping for futility. That is:

$$\sum_{j=1}^{J} Pr_{H_0}\left(\bigcap_{j^*}^{j-1} T_{j^*} \leq c_{E,j^*} \cap T_j > c_{E,j}\right) = \alpha.$$

Notably, there are multiple ways to allocate the type I error rate into different stages, and thus, resulting in various types of stopping boundaries. Pocock (1977) considered the type I error rate allowed for each analysis, α_j, as a constant value smaller than the overall type I error rate α. O'Brien-Fleming (1979) considered very stringent decision thresholds at early stages and the decision threshold decremented with stages. Nowadays, O'Brien-Fleming boundaries are frequently used in evaluating treatment efficacy, because they require the observed treatment effect to be very convincing and they preserve a type I error rate close to that of a fixed-sample design (Chen, Ibrahim, & Chu, 2014).

Other commonly used stopping boundaries include Haybittle-Peto boundaries (Haybittle, 1971; Peto et al., 1976) and boundaries for the triangular test (Whitehead & Stratton, 1983). Haybittle-Peto boundaries assign the same type I error rate to all interim analysis, but consider the type I error allowed at final analysis to be α. The biggest advantage of Haybittle-Peto boundaries is they allow reporting the final results under the pre-defined significance level α. Nevertheless,

the boundaries for interim analyses may be too stringent to terminate a potentially effective treatment. Whitehead and Stratton (1983) also developed stopping boundaries specifically for triangle test. However, the triangle test is less used compared to commonly used classical group-sequential test.

1.2.4 Alpha Spending Function

Notably, a limitation for stopping boundaries discussed above is they require all analyses to be evenly spaced in terms of time or patient enrollment. Therefore, the timing of interim analyses must be specified in advance (O'Brien & Fleming, 1979). To overcome this issue, Lan and DeMets (1994; 1983) proposed alpha spending functions to provide flexibility such that clinical trialists can plan interim analyses at unevenly spaced time points. The alpha spent at each interim analysis is a function of the information fraction, denoted as τ. The information fraction of the *j*th stage, τ_j, is defined as n_j/n_J, where n_J is the maximum sample size required in the clinical trial. Lan and DeMets(1994; 1983) considered four types of alpha spending functions in their papers:

$\alpha_1(\tau) = 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{\tau}\right)$, O'Brien-Fleming

$\alpha_2(\tau) = \alpha ln(1 + (e-1)\tau)$, Pocock

$\alpha_3(\tau) = \alpha\tau^{\theta}$, for $\theta > 0$, Power

$\alpha_4(\tau) = \alpha[(1 - e^{-\gamma\tau})/(1 - e^{-\gamma})]$, for $\gamma \neq 0$.

α_1 and the increments $\alpha_2 - \alpha_1, \dots, \alpha_j - \alpha_{j-1}$ are the type I error rate allowed for the analysis at stage $1,2, \dots, J$. Beta spending functions are developed to allocate the type II error rate into at each analysis in group sequential trials. Similarly, it is shown that beta spending functions control the overall type II error rate at a desired level (Anderson & Clark, 2010).

1.2.5 Frequentist Sample Size Determination for Group-sequential Methods

For classical group sequential clinical trials, the sample size calculation methods have been discussed in the literature (Jennison & Turnbull, 2000; Lehmacher & Wassmer, 1999). Since group sequential methods allow early termination, a maximum sample size (MSS) and an expected sample size (ESS) should be calculated. The maximum sample size is the sample size needed if the trial is not stopping at any interim stages. Since the sample size is proportional to the Fisher information, the maximum sample size can be obtained if the maximum Fisher information is derived. The maximum Fisher information is also the inverse of the variance of $\widehat{\pi_{1J}} - \widehat{\pi_{2J}}$. For original Pocock and O'Brien-Fleming stopping boundaries, Jennison & Turnbull (2000) tabularize a constant ratio of maximum information of a group-sequential clinical trial design to the information of a fixed-sample clinical trial design. The constant ratio is a function of the type I error rate, α, the type II error rate, β, and the number of stages, *J*. For group-sequential design with error spending boundaries, efficacy stopping boundary values are derived recursively for given alpha spent at each stage, if there is an efficacy stopping rule. Then the maximum information is obtained so as to a desired type II error rate (statistical power) is satisfied. Overall, the final maximum sample size can be calculated as:

$$\text{MSS} = n_{fixed} \frac{I_{max}}{I_{fixed}},$$

where n_{fixed} and I_{fixed} are the sample size and Fisher information for the fixed-sample clinical trial design, and I_{max} is the maximum Fisher information for the group-sequential design.

While the expected sample size is the sum of accumulated sample sizes at all stages, each multiplied by the probability of stopping at the corresponding stage. The expected sample size for a clinical trial with two stopping boundaries can be calculated as follow:

$$\text{ESS} = \sum_{j=1}^{J-1} n_j Pr\left(T_j < c_{F,j} \cup T_j > c_{E,j}\right) + n_J Pr\left(\bigcap_{j}^{j-1} c_{F,j} \leq T_j \leq c_{E,j}\right).$$

1.3 Bayesian Methods

Although frequentist methods have dominated clinical trials in the past few decades, they receive some criticisms from clinical trialists. One particular criticism is that the frequentist inference on π is indirect as it calculates the conditional probability of $\hat{\pi}$ given the π. The indirect inference leads to misinterpretations of trial results. A common mistake some clinicians make on interpreting the *p*-value is they think *p*-value is the probability of the null hypothesis to be true (Cohen, 2011). However, the *p*-value does not refer to the probability related to the null hypothesis, actually, the *p* value should be interpreted as the probability to have a treatment effect at least as extreme as the observed value when the null hypothesis is true.

In contrast to frequentist inferences, the other statistical paradigm, Bayesian methods, can assign a probability distribution to π and calculate the probability of π conditional on the observed data (Bayarri & Berger, 2004; D. A. Berry, 2006; J. J. Lee & Chu, 2012). Therefore, clinical trialists are able to assign a probability to the hypothesis by using Bayesian methods. There are some other purported advantages for Bayesian methods, such as incorporating historical information for the treatment effect (D. A. Berry, 2006), as well as fitting naturally with sequential updating in group-sequential methods (Bayarri & Berger, 2004). In the following sections, an overview of Bayesian methods for clinical trials with a specific focus on those for binary outcomes, is presented.

1.3.1 Bayesian Philosophy

Although the establishment of the Bayes theorem by Thomas Bayes can date back as early as the 18th century, Bayesian methodologies were not introduced into clinical trials until 1960 (Anscombe & Aumann, 1963; Cornfield, 1966, 1969; Cornfield & Greenhouse, 1967; Novick & Grizzle, 1965). Anscombe and Aumann (1963) brought up a Bayesian approach to analyze data from group-sequential clinical trials. Novick and Grizzle (1965) proposed a Bayesian probability model for clinical trials with categorical outcomes and discussed the usage of Bayesian priors. Cornfield (1966, 1969, 1976), along with Greenhouse (1967) have made several contributions to implementing Bayesian methods in clinical trials. He advocated using Bayesian approaches to monitor clinical trials. Although frequentist methods dominate the field of clinical trials in the 20th century, Bayesian methods are becoming increasingly popular for clinical trials in recent years. The development of Bayesian methods is largely due to the advances in computer hardware and software, which makes it feasible to compute-intensive algorithms in Bayesian methods, e.g. Markov Chain Monte Carlo sampling methods. The formal support from regulatory agencies also helps popularize Bayesian approaches in clinical trials (FDA, 2014; ICH, 2017). In addition, various standard Bayesian approaches to clinical trials were developed with the improvement of computer power (e.g. D. A. Berry, 2006; S. M. Berry et al., 2010; Freedman & Spiegelhalter, 1989; David J Spiegelhalter, Freedman, & Parmar, 1994).

Philosophically, the Bayesian probability represents the degrees of certainty about a specific event, e.g. the success rate of a new treatment. In the Bayesian paradigm, there is a prior belief about the treatment effect π. The prior belief can be summarized from similar historical studies or be synthesized from the clinician's opionion about the treatment effect. A probability distribution, called prior probability distribution, must be speficied for the prior belief. When data

is collected during the clinical trial, the prior belief about the treatment effect with the new observed evidence from the data. The resulting probability distribution, named the posterior probability distribution, represents the updated knowledge about the treatment effect. The posterior probability distribution is proportional to the product of the prior probability distribution and the likelihood of the observed data, which is also known as the Bayes' theorem, or Bayes' rule.

Another important Bayesian philosophy is that Bayesian methods follow the likelihood principle. The likelihood principle states that all evidence of the treatment effect should come from the data, which ensures the posterior probability distribution of the treatment effect to be appropriate, as long as the prior distribution is specified properly (Diamond & Kaul, 2004; M. D. Lee, 2006).

1.3.2 Bayesian Prior Probability Distribution

As mentioned above, applying Bayesian methods in clinical trials starts with selecting a prior probability distribution. In clinical trials comparing two proportions, the prior distribution of the treatment effect is often specified as a beta distribution:

$$Beta(a_i, b_i), i = 1,2, \tag{1.1}$$

where a and b are the shape parameters of the beta distribution. Such a beta prior takes the advantage of conjugation. To be more specific, with binomially distributed data, the resulting posterior probability also has a beta distribution and is easy to compute. The elicitation of Bayesian beta priors can be intuitive as well, as $a/(a+b)$ can represent the mean or mode for the treatment effect (Fox, 1966; Gross, 1971). Morita et al. (2012) proposed the concept of prior sample size. The authors define $a+b$ as the prior sample size for the clinical trials.

Note, there are several other formal ways to parameterize the beta priors in clinical trials. Thall and Estey (1993) suggested eliciting a beta prior using the prior mean $a/(a+b)$ and the 90 percentile probability interval, denoted as W_{90}, rather than specifying a and b. Such prior specification method is intuitive to clinicians. The authors also recommended to check the tails of the obtained prior distribution to see if the extreme values truly reflect the clinician's opinion on the treatment effect. Ibrahim and Chen (2000) proposed power prior distributions for regression models, which combine heterogenous historical information from multiple previous studies to construct prior distributions.

In addition, Berry and Stangl (1996) used the beta prior distribution assuming parameters are equal, that is $a = b$. A beta prior distribution with equivalent parameters essentially indicates there is no or little prior information about the treatment effect. That is, clinical trialists do not know if the treatment is efficacious or harmful. The $Beta(1,1)$ distribution is commonly used as non-informative prior in clinical trials. This non-informative prior distribution put equal weight to all values of π and maximize the information from data likelihood in the posterior distribution. Hence, Bayesian clinical trials using $Beta(1,1)$ as priors often produce results similar to results of frequentist methods (Lewis, Lipsky, & Berry, 2007). Other non-informative priors appear in the literature include a Jeffery's prior $Beta(0.5, 0.5)$ and a $Beta(0,0)$ prior. Lecoutre and the colleagues (2011) had considered using these two non-informative prior in their researches.

However, some statisticians (Gelman, Simpson, & Betancourt, 2017; Greenland & Poole, 2013) argued that non-informative priors put too much weight on implausibly large values of π. In contrast to non-informative prior, informative priors put changing weights on the values of π, therefore, delivery different information about the treatment effect. The informative priors in clinical trials can be categorized into two classes: optimistic priors and pessimistic priors

(Dersimonian, 1996; Fayers, Ashby, & Parmar, 1997). Optimistic priors (or referred to enthusiastic priors) are a class of priors that consider a beneficial treatment effect and the prior mean is greater than or equal the assumed treatment effect. While pessimistic (or referred as skeptical priors) priors consider a treatment effect is unlikely to be observed, and the prior mean is no greater than the assumed treatment effect. The main reasoning for skeptical priors is the clinical treatment effects are mostly small to moderate, so it may be too unrealistic to have a large treatment effect (Yusuf & Flather, 1995).

Nevertheless, a problem for informative priors is to choose the appropriate prior probability distribution (Hughes, 1993). There are controversial opinions about the use of Bayesian informative priors. On the one hand, people believe prior selection is very subjective and can be easily biased (J. J. Lee & Chu, 2012). Wrongly specified priors can often obscure the results of treatment effects. On the other hand, Fayers et al. (1997) have justified the usage of prior in Bayesian methods. The authors suggested that a pessimistic prior can reduce possibilities to wrongly draw a conclusion that there is overwhelm treatment effect. As such, an optimistic prior can lower the chances that an effective treatment is early terminated for futility.

1.3.3 Bayesian Posterior Probability Distribution

While the prior distribution indicates the prior belief in the treatment effect, the posterior probability shows an updated knowledge about the treatment effect when data is observed. By Bayes' theorem, the posterior probability distribution can be obtained combining data likelihood with respect to the hypothesis and the prior probability distribution. Suppose there are y_{ij} patients having successful outcomes among n_{ij} patients by the *j*th stage, the beta prior distribution **(1.1)** for the *i*th treatment group in is updated to the following posterior distribution:

$$Beta(a_i + y_{ij}, b_i + n_{ij} - y_{ij}), i = 1,2 \quad (1.2)$$

The interpretation of Bayesian posterior probability is intuitive as the posterior distribution directly associate probability with the treatment effect π (J. J. Lee & Chu, 2012). However, the derivation of posterior probability sometimes may require computation intensive methods such as Markov Chain Monte Carlo sampling methods (Brooks, Gelman, Jones, & Meng, 2011).

1.3.4 Bayesian Hypothesis Testing

Based on **(1.2)**, a Bayesian hypothesis test statistic T can be constructed to test the null hypothesis $H_0: \pi_1 - \pi_2 \leq \Delta$ against the alternative hypothesis $H_1: \pi_1 - \pi_2 > \Delta$:

$$T = Pr(\pi_1 - \pi_2 > \Delta), \quad (1.3)$$

where $Pr(.)$ denotes the probability function, π_1 and π_2 each follows the posterior probability distribution in **(1.2)**. When T is greater than some pre-specified decision threshold c, the null hypothesis is rejected. The test statistic **(1.3)** is commonly used in Bayesian clinical trials as it can be nicely interpreted as the probability that the experiment treatment is superior to the control treatment by a minimal clinically significant difference. For example, Xie et al. (2012), Zaslavsky (2012) and Gsponer et al. (2014) used **(1.3)** or its variants in Bayesian hypothesis testing problems.

Kawasaki et al. (2016) proposed a Bayesian equivalent test comparing two proportions which were based on a similar form of the posterior probability. Slightly different from **(1.3)**, the authors defined a two-sided test using the posterior probability, that:

$$T = Pr(-\Delta < \pi_1 - \pi_2 < \Delta), \quad (1.4)$$

where T was also called the κ index by the authors and $-\Delta < \pi_1 - \pi_2 < \Delta$ corresponded to the equivalence hypothesis test in the frequentist paradigm (Shuirmann, 1987). Thus, both **(1.3)** and **(1.4)** were constructed under Bayesian frameworks, but close to test statistics in a frequentist

manner. Although the authors applied the index to analyze the data from a real clinical trial, the authors did not provide a clear guideline on how to make a decision based on the index.

Nonetheless, the calculation of **(1.3)** and **(1.4)** may be difficult because the test statistic does not have a closed-form solution. Monte Carlo sampling techniques are frequently used to estimate the value of these test statistic (Zaslavsky, 2012). The normal approximation can be used to estimate **(1.3)** and **(1.4)** when $\Delta = 0$ (Kawasaki et al., 2016; Zaslavsky, 2012). Remarkably, we developed an efficient algorithm to compute a bulk of values of **(1.3)** with different posterior distributions of π_1 and π_2. The details of the algorithm and improvements on the standard Monte-Carlo sampling procedure can be found in the original paper (Yu, Ramakrishnan, & Meinzer, 2019).

Note, there are other ways to construct the Bayesian test statistic, *T*. Berger and Sellke (1987), Kass and Raftery (1995) and Marden (2000) used Bayes factors to perform the hypothesis test. The Bayes factor is the ratio of the likelihood probability of the alternative hypothesis to the null hypothesis. The larger the Bayes factor is, the stronger the evidence favors the alternative hypothesis. As such, a decision threshold *c* is required to conclude whether to reject the null hypothesis. Jeffreys (1961) provided the original decision threshold value tables for Bayes factors, in which he suggested a Bayes factor greater than 10 to be a strong indicator to reject the null hypothesis. Kass and Raftery (1995) proposed alternative definitions of the decision thresholds, for which the authors suggested a Bayes factor larger than 6 as strong evidence towards the alternative hypothesis.

1.3.5 Bayesian Sample Size Determination

Different from frequentist methods, there is no universal approach to determine the Bayesian sample size for hypothesis testing problems in clinical trials. Instead, there are multiple

ways to estimate the sample size for Bayesian clinical trials, depending on the specific Bayesian hypothesis testing procedure to use. Spiegelhalter (2004) had classified Bayesian sample size determinations into proper Bayesian methods, decision-theoretic Bayesian methods and Bayesian-frequentist hybrid methods. The proper Bayesian' approaches, namely, are a collection of methods to obtain sample size based on pure Bayesian methods. For example, one of the proper Bayesian methods is to estimate the Bayesian sample size by calculating the minimum number of patients required for test statistic **(1.3)** to cross a pre-specified decision threshold *c*. Moreover, sample size determination methods with Bayesian methods as fundamentals can all be classified into proper Bayesian methods (De Santis, 2007; Joseph, du Berger, & Bélisle, 1997; M'Lan, Joseph, & Wolfson, 2008; Weiss, 1997).

Although decision-theoretic sample size determinations are also built based upon Bayesian methods, the class of methods is designated in hypothesis testing under the Bayesian decision-theoretic framework. The decision framework determines optimal sample sizes by maximizing a utility function or minimizing a risk function. A utility function measures several things that clinical trialists are interested in, e.g. the safety of the patients, the effect of the experimental treatment group. A risk function is the expected value of a specific loss function, e.g. mean squared errors. DasGupta and Vidakovic (1997) proposed a sample size calculation method regarding the posterior risk function. The authors sought the minimum sample size that kept the Bayes risk above a threshold value. Katsis and Toman (1999) later extended Gupta and Vidakovic's method by adding a constraint which evaluated the likelihood of the Bayes risk function to exceed the threshold. Sahu and Smith (2006) also proposed an analytic procedure to determine the sample size using the Bayes risk function. The use of decision-theoretical methods, however, is rarely seen in real clinical trials (Müller, Berry, Grieve, & Krams, 2006).

Last, Bayesian-frequentist hybrid approaches combine Bayesian and frequentist methods together to derive the sample size for Bayesian clinical trials. The main reasoning underlying those approaches is to introduce uncertainty into traditional frequentist sample size calculations. An example is the predictive power approach (O'Hagan, Stevens, & Campbell, 2005; Whitehead, Valdés-Márquez, Johnson, & Graham, 2008). The Bayesian predictive power is defined as the expectation of the averaged power, give a specific prior distribution. In the predictive power approach, the power parameter in the classical frequentist sample size calculation formula is replaced by the Bayesian predictive power. The resulting sample size is considered being derived from a hybrid of Bayesian and frequentist methods. Alternatively, Wang et al. (2005) introduced uncertainty to the assumed treatment effect size. The authors replaced the assumed treatment effect size in frequentist sample size calculation with a Bayesian estimate of the treatment effect. Overall, hybrid approaches use Bayesian priors on frequentist methods, while many clinical trialists object to mixing frequentist and Bayesian methods in clinical trials (Inoue et al., 2005).

1.3.6 Bayesian Group-Sequential Methods

Similar to frequentist methods, Bayesian methods can be applied to sequentially analyze clinical trial data. Considerably, Bayesian's philosophy to update the knowledge about π naturally fits the group sequential evaluation (Bayarri & Berger, 2004). Cornfield (1966) first suggested using Bayesian approaches to analyze data from group-sequential trials. A number of Bayesian group designs have been proposed in the past few decades (e.g. Rosner & Berry, 1995; Saville, Connor, Ayers, & Alvarez, 2014; Thall, Simon, & Estey, 1995). Like frequentist methods, specific stopping boundaries need to be determined in Bayesian group-sequential clinical trials.

However, Bayesian stopping boundaries calculations are not as straightforward as those in frequentist methods. Typically, clinical trialists have to use computer simulations to determine the

stopping boundaries that meet certain criteria. For example, Thall et al. (Thall, Simon, & Estey, 1995) constructed the stopping boundaries either to maintain equivalence or to achieve some levels of improvement in the experimental treatment arm compared to the standard of care. Lewis et al. (2007) used simulations to derive stopping boundaries that yielded type I and II error rates similar to frequentist methods. In addition, Gsponer (2013) argued to determine stopping boundaries based on clinicians' opinions. Zhu et al. (2014) adopted the frequentist alpha spending function introduced in section **1.2.4** into Bayesian group sequential clinical trials. However, the authors incorporated only the alpha spending function to determine the Bayesian efficacy stopping boundaries while assuming the futility boundaries to be decided by clinicians and to be constant across all stages.

Currently, there are lots of early phase (i.e. phase I/II) clinical trials adopting Bayesian group-sequential methods (J. J. Lee & Chu, 2012). In comparison, there are less confirmatory clinical trials implementing Bayesian methods. This is because Bayesian models can be very complicated in confirmatory trials with large sample sizes, and may lead to substantial type I error inflation (FDA, 2010). As the Bayesian paradigm does not naturally define the type I error rate, simulations are utilized to demonstrate control of the type I error rate. There is potential simulation bias that clinical trialist may choose specific Bayesian parameters for the null hypothesis to mask the type I error inflation (FDA, 2014; Lai, Lavori, W., & Tsang, 2015). Regulatory agencies are also more conservative and skeptical about the usage of Bayesian informative priors in confirmatory trials (FDA, 2014; ICH, 2017). As a result, there has been a long-time debate between frequentist and Bayesian clinical trialists about whether Bayesian methods should be used in clinical trials in place of classical methods (Efron, 2005; D J Spiegelhalter et al., 1994).

1.4 Reconciliations and Unifications

Despite the long-running methodological debate on frequentist versus Bayesian methods in clinical trials, conflicts between these two methods roots in their philosophies. As discussed in sections **1.2.1** and **1.3.1**, frequentist and Bayesian methods have different definitions for probabilities. Other methodological conflicts in statistical applications between the frequentist and the Bayesian are discussed in the literature (Lindley, 1957; Little, 2006). The most famous one is the Lindley's paradox (1957), where Lindley found that with some specific Bayesian prior distributions, frequentist and Bayesian hypothesis tests could produce opposite results. It is generally accepted these two paradigms are difficult to unify on the philosophical level. However, it is possible to resolve these conflicts on a methodological level (Bayarri & Berger, 2004).

In the following sections, we present a summary of those existing methods. The review starts with methods contributed to general statistics and then narrows down to methods specifically developed for clinical trials. Since frequentist methods are more frequently used than Bayesian methods by the statistical and the clinical trial community for the past few decades, most of the reviewed approaches were developed to reconcile Bayesian methods on frequentist methods. Other literature seeks to resolve disagreements between frequentist methods and Bayesian methods.

1.4.1 Reconciliation Frequentist and Bayesian Methods

Several statisticians have sought correspondences between results of frequentist and Bayesian inferences in statistical data analysis (e.g. J. O. Berger & Sellke, 1987; Casella & Berger, 1987; DeGroot, 1973; Dickey, 1977; Pratt, 1965; Samaniego & Reneau, 1994; Sellke, Bayarri, & Berger, 2007; Lindley, 1965).

Many of the research work attempted to show Bayesian measures of evidence (i.e., posterior probabilities or Bayes factor) consonant with frequentist measures of evidence (i.e., *p*-

values) in different hypothesis testing problems. For several hypothesis testing problems, it was shown that frequentist and Bayesian inferences can produce the same results (e.g. Casella & Berger, 1987). That means the Bayesian posterior probability or Bayes factor is equivalent to the frequentist *p*-value. But in other hypothesis testing problems, Bayesian results seem to be irreconcilable to frequentist results (e.g. J. O. Berger & Sellke, 1987; Dickey, 1977). In general, those reconciliation approaches can be categorized into methods for two-sided hypothesis testing problems and one-sided hypothesis testing problems, Summaries of each approach are provided below.

1.4.2 One-Sided Hypothesis Testing Problems

Lindley (1965) showed that the Bayesian posterior probability distributions of test statistics such as χ^2 or *F* conditional on a large number of observed data can be approximated by the frequentist sampling distribution of the test statistic. Thus, the *p*-value can have a similar interpretation as the tail area of the posterior probability distribution.

DeGroot (1973) further evaluated whether the frequentist *p*-value is compatible with the Bayesian posterior probability of the null hypothesis in tests comparing an arbitrary sample with one or more than arbitrary reference probability distributions. The author defined a class of reference distributions and improper Bayesian priors so that the Bayesian posterior probability of the null hypothesis to be true matched the corresponding frequentist *p*-value.

While DeGroot considered the reconciliation in the context of a relatively broad class of null hypotheses, Casella and Berger (1987) narrowed the class down to composite null hypotheses for the location parameter in one-sided hypothesis tests. Specifically, the authors looked at the null hypothesis $H_0: \pi \leq 0$ versus the alternative hypothesis $H_1: \pi > 0$. Casella and Berger assumed using Bayesian prior distributions that assigned equal mass to both the null and the alternative

hypotheses. In the end, the authors found that for certain reasonable priors, the Bayesian posterior probability that the null hypothesis to be true acquired the same value as the frequentist *p*-value. For other priors, the authors found that resulting Bayesian posterior probabilities encompassed the *p*-value, suggesting the *p*-value might be the lower bound or within the range of Bayesian evidence measures.

Following Casella and Berger, Micheas and Dey (2003) explored the reconciliation between frequentist *p*-values and Bayesian prior and posterior predictive *p*-values using the same setting (i.e. one-sided test for a point null hypothesis for location parameter). The authors considered the null hypothesis $H_0: \pi \leq \pi_0$ versus the alternative hypothesis $H_0: \pi > \pi_0$. The prior and posterior predictive *p*-values are defined as $Pr(Y > y_0 | \pi = \pi_0)$ averaged over the prior and posterior distributions, where Y is the data from a model and y_0 denotes the observed data (Box, 1980; Rubin, 1983). Micheas and Dey found that the infimum of the prior and posterior predictive *p*-values is often the frequentist *p*-value, with a wide class of prior distributions. Later, Micheas and Dey (2007) extended their reconciliation to one-sided hypotheses for scale parameters. Similar results were found when the authors compared frequentist *p*-values and Bayesian prior and posterior predictive *p*-values for a number of priors. For many prior specifications, the infimum of Bayesian predictive *p*-values was equal to the classical frequentist *p*-value.

Alternatively, Yin and the colleagues measured the reconcilability between frequentist and Bayesian methods in hypothesis tests with the presence of nuisance parameters (Yin, 2011; Yin & Wang, 2016; Yin & Zhao, 2013). Nuisance parameters are those parameters which are not of primary interest but still should be included in the statistical model when estimating the main effect. Inspired by Casella and Berger (1987), Yin and colleagues considered the same problem of testing one-sided hypotheses. However, the authors checked the reconcilability between the frequentist *p*-

value and the Bayesian posterior probability with additional assumptions, e.g. tests for a scale parameter under an exponential distribution (Yin, 2011), test for normal means (Yin & Zhao, 2013) and tests location-scale parameters under a Weibull distribution (Yin & Wang, 2016). Results from Yin and the colleagues' work suggested there was still a match between the frequentist *p*-value and Bayesian posterior probability under certain circumstances, in consideration of nuisance parameters.

Altham (1969) considered the reconciliation problem within a slightly different context. The author investigated the relationship between Fisher's exact probability and the Bayesian posterior probability resulted from analyzing a 2 x 2 contingency table. Altham proved that there was an equivalence between exact probability and the posterior probability by using the identity between the incomplete beta function and the cumulative binomial distribution. The identity was mentioned earlier in publications of Hartley and Fitch (1951) and Raiffa and Schlaifer (1961). Zaslavsky and the colleagues (Zaslavsky, 2010, 2012; Zaslavsky & Scott, 2012) drew a similar conclusion on the relationship between Fisher's exact probability and the Bayesian posterior probability. Differently, Altham calculated the Bayesian posterior probability in terms of the odds ratio, while Zaslavsky considered the Bayesian posterior probability in terms of the success rate difference.

1.4.3 Two-Sided Hypothesis Testing Problems

Although reconciliations between frequentist and Bayesian evidence of measure for one-sided hypothesis testing problems seemed feasible when specific priors were chosen, reconciliations for two-sided hypotheses were not always successful.

Dickey (1977) considered the reconciliation between the frequentist *p*-value and the Bayes factor for the hypothesis testing problem with a point null hypothesis $H_0: \pi = 0$ versus a two-sided

composite alternative hypothesis $H_1: \pi \neq 0$. The author made an assumption that observed sample data were from a normal distribution. Furthermore, the Bayes factor is defined as the posterior distribution to the prior distribution, evaluated at H_1 for $\pi = 0$. At last, Dickey found that the frequentist *p*-value was consistently smaller than the infimum of the Bayes factor.

Berger and Selke (1987) also considered reconciling frequentist and Bayesian measures of evidence in two-sample two-sided hypothesis testing problems with normal endpoints. The same hypotheses as in Dickey (1977), $H_0: \pi = 0$ versus $H_1: \pi \neq 0$ was tested. However, the authors evaluated the Bayesian posterior probability of the null hypothesis instead of the Bayes factor. The posterior probability conditional on the observed data y is defined as $P(H_0|y)$. The frequentist *p*-value was calculated based on the same data. Berger and Selke found that regardless of which prior distribution was used, Bayesian approaches always yielded posterior probabilities above the pre-defined decision threshold, leading to fail to reject the null hypothesis. On the contradictory, the calculated frequentist *p*-values were always less than the significance level of 0.05, which meant the null hypothesis should be rejected. Hence, a conclusion could be made that frequentist *p*-values and Bayesian posterior probability of the null hypothesis were irreconcilable when analyzing a two-sample two-sided hypothesis testing problem assuming normal endpoints.

Nonetheless, De La Horra (Horra, 2005) showed that it was possible to reconcile classical and prior predictive *p*-values in the two-sided hypothesis testing problems for location parameter. Following Micheas and Dey's reconciliation work for location parameter on one-sided hypothesis tests (Micheas & Dey, 2003), De La Horra computed the prior predictive *p*-value by averaged $Pr(Y > y_0 \; or \; Y < -y_0 | \pi = \pi_0), y_0 > 0$ over the prior distribution. The author found that for some given prior distributions, the frequentist *p* value is equivalent to the Bayesian prior predictive *p*-value.

In addition, Dempster (1973), Aitkin (1997) and Smith and Ferrari (2014) explored the correspondence between frequentist *p*-value and the posterior in likelihood ratio test. Dempster (1973) proposed the posterior distribution of likelihood ratio (PLR) for hypothesis testing problems in the simple versus two-sided composite case. The PLR was defined as $Pr(LR(\pi, y) \leq \zeta | y)$, where $LR(\pi, y)$ was the ratio of the likelihood evaluated at the null to the maximum likelihood evaluated at the alternative, and ζ was a decision threshold for the likelihood ratio test. Dempster further accessed whether the PLR might match the frequentist *p* value of the likelihood ratio test assuming data was from a normal distribution. Finally, the author found that when a uniform prior and $\zeta = 1$ were applied, the PLR equaled th *p* value.

The same comparison between the PLR and the *p*-value was carried out by Aitkin (1997), and similar conclusions were reached. However, Aitkin (1997) extended Dempster's (1973) PLR definition to more general distributions. But, as long as a smooth prior distribution and $\zeta = 1$ were applied, the PLR asymptotically matched the the *p* value. Smith and Ferrari (2014) further made the reconciliation results between the PLR and the *p* value more generalized, to an invariat framework. The invariance property allowed establishing an equivalence between frequentist confidence intervals and Bayesian credible intervals.

1.4.4 Fixed-Sample Clinical Trials

While for general statistic, the literature review suggests that lots of efforts have been made to reconcile frequentist and Bayesian measure of evidence during statistical inferences. Successful reconciliations often require specific assumptions related to the hypotheses, endpoints and prior selection. Compared to two-sided hypothesis testing problems, one-sided hypothesis testing problems were easier for frequentist *p*-value and Bayesian posterior probability, or Bayes factor to achieve reconciliation. For clinical trials, several methods seeking the parallel between

frequentist and Bayesian methods have also been proposed. However, a large proportion of these methods are more focused at the design stage rather than the analysis or the inference stage. In addition, a lot of the reconciliation methods for the frequentist *p*-value and the Bayesian measure of evidence are not applicable to clinical trials. This is because post-adjustments on Bayesian prior distributions may be inappropriate and clinical trials it is more of interest to achieve the same type I or II error rates between frequentist and Bayesian approaches, as type I and II error rates are required to be strictly controlled (FDA, 2014). As stated near the end of section **1.1**, we aimed to develop a novel unified approach for frequentist and Bayesian methods for clinical trials with binary endpoints. The following sections are concerned with existing reconciliations or unifications between frequentist and Bayesian methods in clinical trials. Details and limitations of each method are provided below. Since most Bayesian methods for clinical trials use one-sided hypothesis test and unification methods are developed either assumed there was only one stage or multiple stages, the related literature is categorized into methods for fixed-sample clinical trials and group-sequential clinical trials.

Inoue, Berry, Parmigiani (2005) showed that frequentist and Bayesian methods can yield the same sample size at the design stage in single-arm clinical trials. The authors' motivation came from the commonality between frequentist and Bayesian sample size determinations that both methods sought the minimum sample size to achieve certain goals. Inoue, Berry, and Parmigiani first clearly defined the frequentist and Bayesian goal functions with respect to the sample sizes, respectively. Let subscripts *F* and *B* denote the frequentist and Bayesian parameter. The frequentist goal was to reach the power $1-\beta$ for an assumed treatment effect, meanwhile controlling the type I error α at the desired level, and the sample size can be defined as:

$$n_F = min\{n \in \mathbb{N}, G_F(n, \boldsymbol{u}) > G_F^*\},$$

where $\mathbb{N}$ denoted the set of integers, $G_F(n, \boldsymbol{u})$ denoted the frequentist goal function with a unique frequentist inputs $\boldsymbol{u}$ and the G_F^* was the desired level of the goal function and was held as a constant. While for Bayesian sample size calculation, the goal was to reach a targeted rate, named classification rate, for correctly making the decision whether to reject the null hypothesis conditional on prior distributions:

$$n_B = min\{n \in \mathbb{N}, G_B(n, \boldsymbol{v}) > G_B^*\},$$

where $\boldsymbol{v}$ are Bayesian inputs for the goal function. Although goal functions were quite different for frequentist and Bayesian methods, the authors were able to reach a unification between frequentist and Bayesian sample sizes by setting $n_F = n_B$. Moreover, the authors identified a unique mapping between the assumed treatment effect in frequentist sample size calculation and the prior variance in Bayesian sample size calculation. To be more specific, the authors calculated the frequentist sample size for a given treatment effect and calibrated the variance of the prior distribution to produce the same sample size.

The unification method developed by Inoue, Berry, Parmigiani (2005) had great significance because it unified frequentist and Bayesian methods at the design phase of clinical trials. As clinical trials required to analyze the data as planned, it was inappropriate to reconcile the frequentist and the Bayesian during data analysis. In addition, the mapping between frequentist and Bayesian methods on the sample size allowed clinical trialists to conduct clinical trials based on their own preference.

However, there were several limitations to the unification method. First, it only considered single-arm clinical trials, while most modern clinical trials are designed to compare two or more treatments. Second, the Bayesian classification rate was not the same as the frequentist type I error

rate and is difficult to interpret it to clinicians. It could be difficult for a clinical trial using Bayesian classification rate while ignoring type I error rate to get approved from regulatory agencies as well. For instance, in the paper, the authors derived the same sample size between frequentist and Bayesian approaches when setting the frequentist assumed treatment effect to be 0.1, type I error rate to be 0.05 and the Bayesian classification rate to 0.9283. The meaning of 0.9283 is relatively hard to interpret and the classification rate did not guarantee the type I error rate of the Bayesian approach was also controlled at 0.05. In other words, the frequentist and Bayesian methods were unified at the design stage for the sample size, but not necessarily in the analysis stage for the same type I or type II error rates or conclusion.

Zaslavsky (2010, 2012) considered reconciliation between frequentist and Bayesian measures of evidence methods for several types of fixed-sample clinical trials. The author restricted the hypothesis tests to one-sided superiority or non-inferiority tests. Further, the author assumed using conjugate beta distribution with integer parameters as priors and $T = Pr(\pi_1 - \pi_2 > \Delta)$ as the Bayesian statistic for hypothesis tests. Zaslavsky (2010) compared the Fisher's exact *p*-value with Bayesian posterior probability under single-arm fixed-sample clinical trials with binary and Poisson outcomes. The author found the *p*-value can be smaller, equal or greater than the posterior probability for different choices of prior distributions.

Later, Zaslavsky (2012) showed frequentist and Bayesian hypothesis testing results were reconcilable in two-arm fixed-sample clinical trials with binary endpoints. Under these conditions, Zaslavsky proved that a Fisher's exact *p*-value can transform into the Bayesian posterior probability by adjusting the number of events *y* and total sample size *n*:

$$Pr_F\left(\pi_1 - \pi_2 > \Delta | y_1 + a_1, n_1 + a_1 + b_1 - 1, y_2 + a_2 - 1, n_2 + a_2 + b_2 - 1\right)$$

$$= Pr_B\left(\pi_1 - \pi_2 > \Delta | y_1, n_1, y_2, n_2\right) + f(\Delta),$$

where $f(\Delta)$ is a well-defined function of Δ by the author. By using this relationship, Zaslavsky found that *p*-values of Fisher's exact test and Bayesian posterior probabilities were close when non-informative priors were used. Note, the relationship between *p*-values and posterior probabilities above was very similar to that established by Altham (1968) and Zaslavsky (2010). Here, the major difference was Zaslavsky (2013) extended the relationship to a more general null hypothesis, to include the minimally clinically significant difference, Δ.

Publications of Zaslavsky's research (Zaslavsky, 2010, 2012) had great importance because they showed connections between frequentist and Bayesian methods in clinical trials and proposed methods had potential usage in the comparison between frequentist and Bayesian methods. Biostatisticians were capable to use frequentist and Bayesian calculations interchangeably. For example, when the sample size was small, the Bayesian operating characteristics can be determined by using available software for frequentist methods, while for large sample size, frequentist characteristics can be computed using Bayesian simulations. However, it should be noticed that Zaslavsky's approaches assumed exact test, non-informative priors with integer parameters, which restricted the use in planning clinical trials. Also, the same p-values and posterior probabilities did not mean the final type I and II error rates were the same for frequentist methods and Bayesian methods. Thus, these approaches were less useful in designing practical clinical trials.

In addition, Zaslavsky and Scott (2012) took a relatively different direction and worked on the reconciliation between frequentist and Bayesian methods for comparing the mean of multiple

single samples with a constant value. The authors set the problem for continuous and discrete outcomes that were commonly seen in clinical trials, e.g. normal or binary endpoints. Data from both distributions could be evaluated using a normally distributed test. Zaslavsky and Scott proposed a unified procedure of hypothesis testing problems involving adjustments for multiple testing, for frequentist and Bayesian methods. The unified procedure calculated order statistics using adjusted interval limits. The interval limits were defined as the confidence interval for frequentist methods and the credible interval for Bayesian methods, respectively. The modification on credible limits was similar to that on the confidence limits because the authors adjusted the credible limits with respect to an established concept of reconciling the frequentist *p*-value and the Bayesian posterior probability (Zaslavsky, 2010).

Results from a numerical study indicated the family-wise type I error rate was controlled for both frequentist and Bayesian methods by using the unified procedure. However, to use order statistic for multiple testing is less popular than using other adjustments such as the Bonferroni correction. The authors also mentioned that computation could be an issue since the convergence of approximate estimates for the required sample size sometimes could be slow.

1.4.5 Group-Sequential Clinical Trials

Compared to methods for fixed-sample clinical trials which might focus more on the theoretical aspects, the reconciliations and unification between frequentist and Bayesian methods in group-sequential clinical trials concentrated on the application aspects. Though clinical trialists seldom use frequentist methods and Bayesian methods together except for that FDA requires to re-assess a Bayesian confirmatory trial using frequentist test procedures, the possibility to combine frequentist characteristics and Bayesian methods in the same clinical trial is studied in the literature. For most of the unification methods in the literature, the authors wanted to control the frequentist

type I error rate, a concept to be strictly controlled as required by regulatory agencies, while using Bayesian methods.

Lewis et al. (2007) developed a Bayesian decision-theoretic framework for two-arm group-sequential clinical trials. The core of the decision-theoretic framework was a Bayesian utility function. The utility function measured several components including the benefits of treatment and the cost of future patient enrollment. An optimal Bayesian design was attained when the value of the utility function was maximized.

When Lewis et al. configured the utility function to produce desired frequentist characteristics, results showed Bayesian methods required a smaller sample size compared to the frequentist methods. The main reason was that Bayesian clinical trials had more frequent interim looks than frequentist methods. The authors found that the sample sizes of frequentist and Bayesian approaches were almost the same when setting the number of analyses for two methods to be the same and assuming a non-informative prior for Bayesian approaches. In addition to the sample size, the type I error rates were the same for frequentist and Bayesian methods when the non-informative prior was applied.

Similarly, Ventz and Trippa (2014) reconciled frequentist methods and Bayesian methods in multi-arm multi-stage clinical trials within the Bayesian decision-theoretic framework. The authors defined the utility function to represent the cost of infrastructure in clinical trials and enrolled patients at each stage, and let the utility function to satisfy the frequentist characteristics with a constraint. As such, the optimal Bayesian design was obtained by maximizing the expected value of the utility function. The resulting Bayesian design preserved the optimal Bayesian characteristics and a Bayesian interpretation, while controlled the frequentist type I error rates.

Extensive numerical studies verified that the overall type I error rate was controlled as desired, for different specifications of the utility function.

Although both Lewis et al. (2007) and Ventz and Trippa (2014) developed approaches to reconcile frequentist methods and Bayesian methods in clinical trials, it is relatively difficult to implement these approaches in real clinical trials. First, the usage of the decision-theoretical framework in clinical trials is rare. The decision-theoretical framework was less accepted and understood by regulatory agencies compared to Bayesian hypothesis testing based on the posterior probability (FDA, 2014). Second, the determination of the utility function can be very subjective and to interpret each element in the utility function and explain how each element could affect the final results can be confusing. Last, finding the optimal design often requires numerous simulations. The computation becomes intractable when the sample size is large.

Notably, some other Bayesian methods than decision-theoretical approaches were developed in order to control the frequentist type I error. Zhu (2015) incorporated the frequentist alpha spending functions into Bayesian group-sequential designs. The authors proposed a simulation-based algorithm to derive specific efficacy stopping rules in Bayesian group-sequential designs. The algorithm first simulated a large number of samples from the posterior distribution of the statistic **(1.3)**, assuming $\Delta = 0$. The efficacy boundary for the *j*th analysis is then determined as the $\left(1 - \alpha\left(\tau_j\right)\right) * 100\%$th percentile of the ordered simulated samples. The final results indicated that the frequentist type I error rate was well controlled at the desired level. Nonetheless, Zhu et al. did not consider introducing a corresponding beta spending function for futility boundaries. Instead, the authors made a claim that the futility boundaries were the same across all analyses and were determined based upon opinions of clinicians.

Zhu et al. (2015) utilized a hybrid approach which consists of frequentist concepts and a Bayesian framework. The design utilized alpha-spending function and the results showed unification on type I error rate was achieved between frequentist and Bayesian group-sequential clinical trials in some scenarios. While for other scenarios, Bayesian type I error rates were smaller than the frequentist values. There were several noticeable limitations. First, the approach requires obtaining the posterior probability distribution using simulations, which could be time-consuming and generate simulation bias. The simulation bias could affect the efficacy boundaries. Second, the correlation between Bayesian posterior probabilities across stages was not considered in the simulation-based approach, which might lead to wrong alpha spent at each stage for the Bayesian methods. Third, prior distributions were not considered when computing the Bayesian efficacy boundaries. Finally, the way to specify futility boundaries and control them as fixed values are not common cases in clinical trials. Type II error rate was not well controlled and might inflate in the Bayesian methods.

Shi and Yin(Shi & Yin, 2019) proposed another method to control type I error rate in single boundary Bayesian group sequential trials. The method reduced the computation burden required by simulations and maintained the desired type I error rate. However, the authors did not consider stopping the trial for futility at the interim analysis. Although both Zhu et al. and Shi and Yin reconciled frequentist methods to Bayesian methods to some extent, the final results were sensitive to Bayesian prior specifications. To be more specific, the Bayesian type I error rate can vary because stopping boundaries are computed without taking prior distributions into account. Furthermore, the other frequentist concept, the type II error rate, may not be preserved as desired either, under different prior distributions.

1.5 Appendix

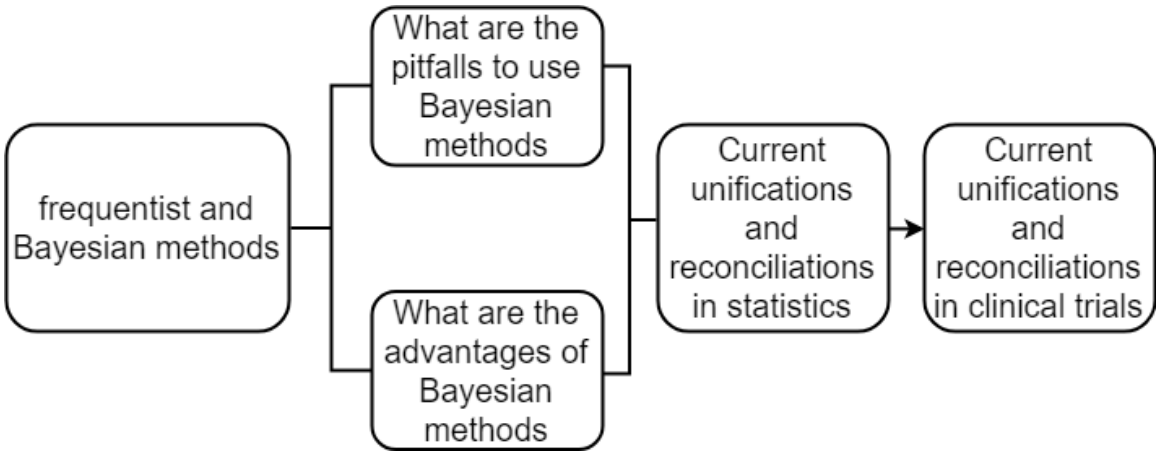


Figure 1.1 The workflow of the literature review. The review starts with understanding frequentist and Bayesian methods and their fundamental conflicts. Then the advantages and disadvantages of the Bayesian approach are reviewed to help crystalize our specific aims. Finally, existing methods on unifying frequentist and Bayesian methods are reviewed.

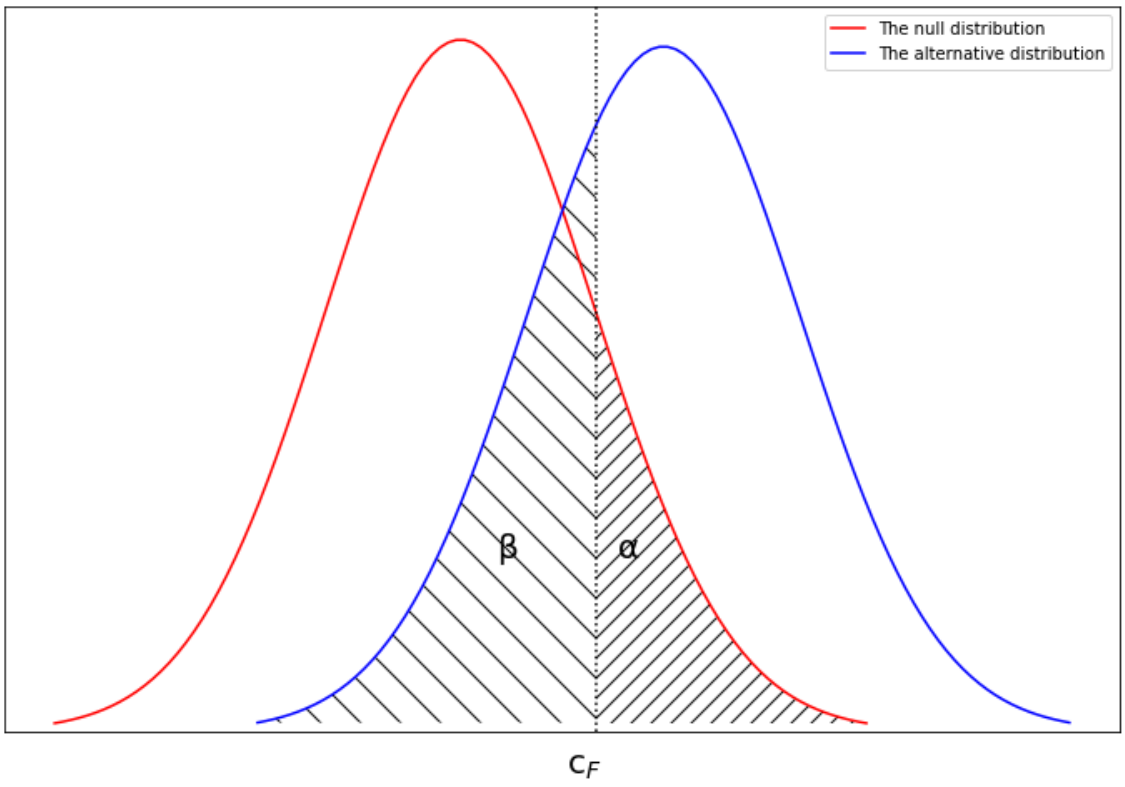


Figure 1.2 Relationships among α, β and c, distributions of the test statistic T under the null (the red curve) and under the alternative (the blue curve) in frequentist hypothesis testing. The c is determined as the threshold that the area to the right tail of the null distribution is α and the area to the left tail of the alternative distribution is β.

CHAPTER TWO: SPECIFIC AIMS

A major limitation spotted from the literature review, is that there is no theoretical method in clinical trials to achieve the same type I and II error rates between frequentist and Bayesian methods. Moreover, it should be noticed that many methods in the literature review require modifications on Bayesian priors to achieve the reconciliation or unification between two approaches. Nevertheless, it becomes inappropriate to make post adjustments on Bayesian priors when conducting clinical trial data analysis. In fact, any modification on the prior distribution can be inappropriate for clinical trials, because prior distribution is often specified to reflect the prior belief in treatment effect. Thus, novel unified approaches are needed to establish a correspondence between frequentist and Bayesian methods on type I and II error rates, without changing a specific Bayesian prior distribution. Since the Bayesian prior distribution is designated to be held fixed, it is possible to adjust the sample size and decision threshold for Bayesian methods to achieve the unification.

In addition, some other limitations occurred in the literature for reconciliations or unifications between frequentist methods and Bayesian methods are summarized. Some previous work used very complicated analytic forms or simulations, which should cost a lot of time to implement (Shi & Yin, 2019; Zaslavsky, 2012; Zhu & Yu, 2015); some methods were conceptually complicated or were difficult to apply in clinical trials (Inoue et al., 2005; Lewis et al., 2007; Ventz & Trippa, 2014); some results were not generalized enough, e.g. some methods was only applicable to single boundaries (Shi & Yin, 2019; Zhu & Yu, 2015). We also aim to address these limitations in our proposed approaches. Thereby, the following three specific aims are proposed. Reasonings underlying three aims are illustrated in **Figure 2.1**.

2.1 Aim 1. A unified approach for frequentist and Bayesian methods in two-arm fixed sample clinical trials with binary endpoints.

First, the distribution of the Bayesian posterior probability will be derived. Second, a one-to-one mapping between frequentist and Bayesian methods will be established by utilizing the straightforward classical frequentist hypothesis testing framework. The one-to-one mapping will provide unified type I and II error rates between frequentist and Bayesian methods. Third, a theoretical approach to determine the sample size and decision threshold for the Bayesian approach will be developed. Finally, a comparison is made between two approaches on type I and II error rate, the sample size will be made under different scenarios through a numerical study.

2.2 Aim 2. A unified approach for frequentist and Bayesian methods in group sequential clinical trials with binary endpoints.

Aim 2 is concerned with harmonizing and frequentist methods in group sequential clinical trials. The distribution of posterior probability in **Aim 1** will be extended to a multivariate case. Based on the multivariate distribution, theoretical approaches to calculate corresponding sample size and stopping boundaries will be provided. Detailed algorithms will be illustrated. Similarly, a comparison will be carried out between frequentist methods and Bayesian methods on type I and II error rates, the maximum sample size, the expected sample sizes and stopping boundaries.

In traditional group sequential designs, alpha spending function and beta spending function are introduced at each analysis to achieve the desired overall type I error rate. **Aim 2** is also concerned with applying alpha spending function and beta spending function in Bayesian methods, and achieve the same type I and II error rate for each analysis, as corresponding frequentist methods.

2.3 Aim 3. Implementation of the unification framework.

Aim 3 is concerned with practical issues to implement proposed unification approaches in **Aim 1** and **Aim 2**. First, a complete investigation will be conducted to evaluate the influence of Bayesian prior specifications, allocation ratios, and numbers of analyses, on the final sample sizes. The results will then be compared to frequentist results, and a suggestion for choosing between frequentist methods and Bayesian methods will be given. Second, a user-friendly software application to implement the proposed unified approaches will be deployed online, so that other clinical trialists can use our approaches without writing any code.

2.4 Appendix

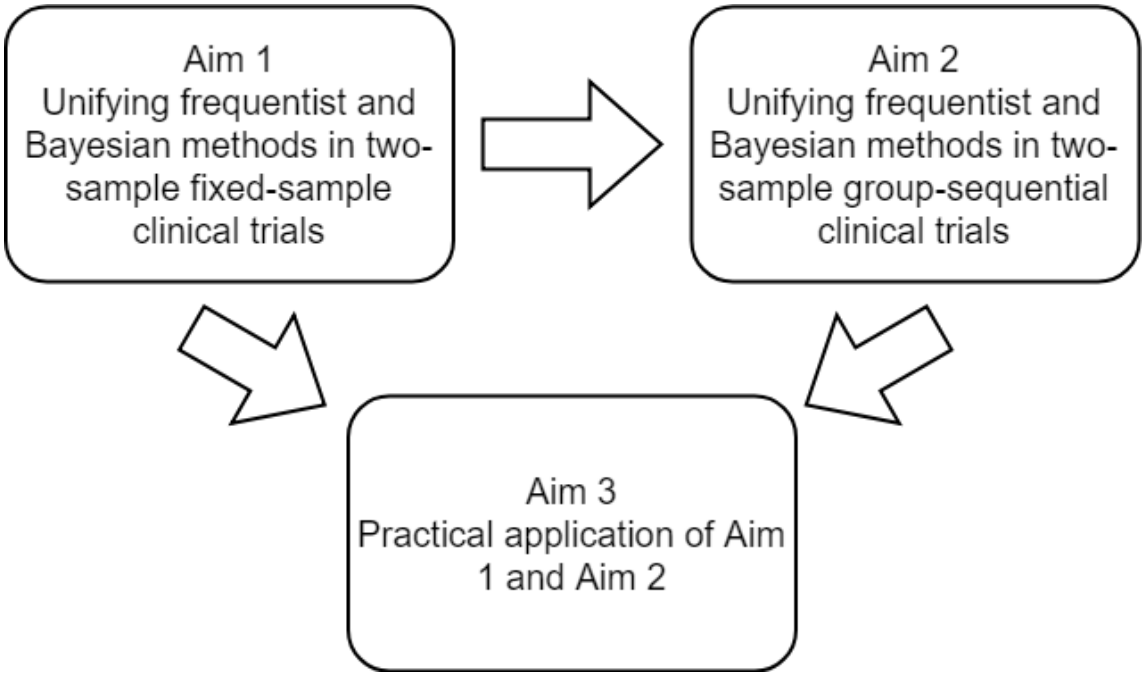


Figure 2.1 Linear progression of logic for three specific aims. The **Aim 1** proposes a unified approach for frequentist and Bayesian methods applied to the simplest case in a clinical trial, which is the fixed-sample clinical trials. Then the **Aim 2** improves the specific **Aim 1** and extends the methodologies to group sequential clinical trials, which is more commonly seen nowadays. Finally, **Aim 3** considers the practical implementation and usage of two unified approaches proposed in **Aim 1** and **Aim 2**.

CHAPTER THREE: SPECIFIC AIM 1

3.1 Introduction

In this chapter, a novel unified approach is proposed for frequentist and Bayesian methods in one-sided two-arm fixed-sample clinical trials with binary endpoints. It is assumed conjugate beta priors are used for Bayesian methods. The novelty in the proposed approach is, it unifies frequentist and Bayesian methods by generating not only the same type I error rate, but also the same type II error rate in the analysis. In **Section 3.2** how the unification is achieved between the frequentist and Bayesian approaches is discussed. A theoretical argument for all the claims made is also provided. **Sections 3.2** and **3.3** present a one-to-one correspondence between frequentist and Bayesian designs in clinical trials with binary endpoints.

3.2 Methods

To compare the experimental arm and the control arm in a clinical trial, considering a similar setting as introduced in **Chapter 1**. The test of the following one-sided hypothesis is of interest: $H_0: \pi_1 - \pi_2 \leq \Delta$ versus $H_1: \pi_1 - \pi_2 > \Delta$ where π_1 and π_2 are success rates of the experimental arm and the control arm respectively, and Δ is the minimal clinically important difference. Here, Δ is defined as the minimum change in outcomes to prove the experimental arm is clinically superior to the control arm. In order to conduct hypothesis tests, consider two elements introduced in **Chapter 1 Section 1.2.2**, the statistic T and the pre-specified decision threshold c. The evidence of H_0 is rejected in favor of H_1 when $T > c$. The details of frequentist and Bayesian hypothesis tests have been illustrated in **Section 1.2** and **1.3**, only a few notations should be emphasized here. A conjugated beta prior distribution for treatment i is specified, $\pi_i \sim Beta(a_i, b_i), i = 1,2$, where a_i and b_i are shape parameters of the distribution. The uncertainty about π_i is updated when y_i

successful events are observed for treatment *i* in the clinical trial. A posterior probability distribution is used to represent the updated uncertainty, $\pi_i \sim Beta(a_i + y_i, b_i + n_i - y_i)$.

Further, let subscript *F* and *B* to denote parameters specifically for frequentist methods and Bayesian methods. The Bayesian statistic T_B is constructed as the posterior probability that π_1 is superior to π_2 by at least Δ. That is, $T_B = \Pr(\pi_1 - \pi_2 > \Delta)$, which is the same as **(1.3)**. Here, the form of the posterior probability is chosen because it is intuitive to interpret. For example, the result of $\Pr(\pi_1 - \pi_2 > 0.05) = 0.95$ can be interpreted as "the probability of the experimental arm is better than the control arm by at least 0.05 is 95%". The corresponding type I and II error rates for the Bayesian approach are defined as follows and can be simulated using Monte Carlo sampling techniques:

$$\alpha_B = Pr(Pr(\pi_1 - \pi_2 > \Delta) > c_B | H_0),$$

$$\beta_B = Pr(Pr(\pi_1 - \pi_2 > \Delta) \leq c_B | H_1).$$

Bayesian methods do not necessarily lead to the same error probabilities as frequentist methods(FDA, 2014). Nevertheless, correspondences still can be established between frequentist and Bayesian methods on error probabilities. By definition, there is a tradeoff between the type I error rate, α and the type II error rate, β when the total sample size n is held fixed. For instance, if the value of c_B increases, α_B should decrease while β_B should increase, where α_B and β_B denote the Bayesian type I and II error rates. Therefore, we propose a unified approach for frequentist and Bayesian hypothesis testing to produce the same error rates based on adjustments of Bayesian threshold c_B and Bayesian sample size n_B given that $\alpha_B = \alpha_F$ and $\beta_B = \beta_F$. Note, while previous methods have shown that it is possible to adjust priors and Δ for the Bayesian tests to achieve the unification, we are reluctant to do so because priors and Δ are often dictated as pre-specified

quantities of clinical interest in clinical trials, and thus, modification of these parameters can influence basic clinical assumptions.

Consider the relationships among α_F, β_F and c_F discussed by Lachin (Lachin, 1981) in frequentist methods in **Figure 1.1**, similar relationships among α_B, β_B and c_B also can be identified if the null and alternative distributions of $\Pr(\pi_1 - \pi_2 > \Delta)$ are known. However, the distribution of $\Pr(\pi_1 - \pi_2 > \Delta)$ is highly skewed and does not have a closed-form solution. Thus, for the purposes of deriving a closed form solution that does not rely on simulation, a standard normal quantile transformation on $\Pr(\pi_1 - \pi_2 > \Delta)$ is applied and the resulting distribution can be well approximated by the normal distribution when n_B is sufficiently large.

Although, the method discussed here would hold for any continuous distribution, for convenience a normal approximation is considered. However, in most clinical trial scenarios, especially for those Phase III confirmatory trials, where prior information is realistically available and controlling frequentist characteristics is required by regulatory agencies assumption of normality has been found adequate. Let η denote the transformed variable, that is, $\eta = \Phi^{-1}\big(\Pr(\pi_1 - \pi_2 > \Delta)\big)$, where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution. Still, let P_1 and P_2 denote the assumed treatment effects for the experimental arm and control arm, and $\widehat{\pi_1}$ and $\widehat{\pi_2}$ denote the treatment effect estimates for the experimental arm and the control arm, where $\widehat{\pi_1} = y_1/n_1$ and $\widehat{\pi_2} = y_2/n_2$. Let $Beta(a_1, b_1)$ denote the prior for the experimental arm and $Beta(a_2, b_2)$ denote the prior for the control arm. The distribution of η under H_0 can be approximated using the following normal distribution:

$$\eta|H_0 \sim N\left(\frac{\mu_1 - \mu_2 - \Delta}{\sqrt{\frac{\mu_1(1-\mu_1)}{n_1 + a_1 + b_1 + 1} + \frac{\mu_2(1-\mu_2)}{n_2 + a_2 + b_2 + 1}}}, \frac{\frac{n_1 P_2(1-P_2)}{(n_1 + a_1 + b_1)^2} + \frac{n_2 P_2(1-P_2)}{(n_2 + a_2 + b_2)^2}}{\frac{\mu_1(1-\mu_1)}{n_1 + a_1 + b_1 + 1} + \frac{\mu_2(1-\mu_2)}{n_2 + a_2 + b_2 + 1}}\right), \qquad (3.1)$$

where μ_1 and μ_2 are the hypothesized posterior mean for the experimental arm and the control arm respectively, and $\mu_1 = (n_1P_2 + a_1)/(n_1 + a_1 + b_1)$ and $\mu_2 = (n_2P_2 + a_2)/(n_2 + a_2 + b_2)$.

Similarly, the distribution of η under H_1 can be approximated using the following normal distribution, where $\mu_\cdot^*$ denotes the hypothesized posterior mean under the alternative.

$$\eta|H_1 \sim N\left(\frac{\mu_1^* - \mu_2^* - \Delta}{\sqrt{\frac{\mu_1^*(1-\mu_1^*)}{n_1 + a_1 + b_1 + 1} + \frac{\mu_2^*(1-\mu_2^*)}{n_2 + a_2 + b_2 + 1}}}, \frac{\frac{n_1P_1(1-P_1)}{(n_1 + a_1 + b_1)^2} + \frac{n_2P_2(1-P_2)}{(n_2 + a_2 + b_2)^2}}{\frac{\mu_1^*(1-\mu_1^*)}{n_1 + a_1 + b_1 + 1} + \frac{\mu_2^*(1-\mu_2^*)}{n_2 + a_2 + b_2 + 1}}\right), \quad (3.2)$$

where $\mu_1^* = (n_1P_1 + a_1)/(n_1 + a_1 + b_1)$ and $\mu_2^* = (n_2P_2 + a_2)/(n_2 + a_2 + b_2)$. The Proof of the distribution of η is shown blow:

Proof 3.1

Kawasaki et al. (Kawasaki et al., 2016) have shown that $Pr(\pi_1 - \pi_2 > \Delta)$ can be approximated by the CDF of a standard normal distribution. That is,

$$Pr(\pi_1 - \pi_2 > \Delta) \approx \Phi\left(\frac{\mu_{Y1} - \mu_{Y2} - \Delta}{\sqrt{\frac{\mu_{Y1}(1-\mu_{Y1})}{n_1 + a_1 + b_1 + 1} + \frac{\mu_{Y2}(1-\mu_{Y2})}{n_2 + a_2 + b_2 + 1}}}\right),$$

where μ_1 and μ_2 are posterior means of π_1 and π_2, and $\mu_{y_1} = (y_1 + a_1)/(n_1 + a_1 + b_1)$ and$\mu_{y_2} = (y_2 + a_2)/(n_2 + a_2 + b_2)$. Here, y_1 and y_2 are two independent random variables and follow binomial distributions that $y_i \sim Binomial(n_i, P_i), i = 1,2$. When *n* is large, we have that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} n_1P_1 \\ n_2P_2 \end{bmatrix}, \begin{bmatrix} n_1P_1(1-P_1) & 0 \\ 0 & n_2P_2(1-P_2) \end{bmatrix}\right).$$

Therefore, applying the delta method, the asymptotic distribution of $\eta = \Phi^{-1}\big(\Pr(\pi_1 - \pi_2 > \Delta)\big)$ is a normal distribution, that

$$\eta \sim N\left(\frac{\mu_1-\mu_2-\Delta}{\sqrt{\frac{\mu_1(1-\mu_1)}{n_1+a_1+b_1+1}+\frac{\mu_2(1-\mu_2)}{n_2+a_2+b_2+1}}}, \frac{\frac{n_1P_1(1-P_1)}{(n_1+a_1+b_1)^2}+\frac{n_2P_2(1-P_2)}{(n_2+a_2+b_2)^2}}{\frac{\mu_1(1-\mu_1)}{n_1+a_1+b_1+1}+\frac{\mu_2(1-\mu_2)}{n_2+a_2+b_2+1}}\right),$$

where $\mu_1 = (n_1P_1 + a_1)/(n_1 + a_1 + b_1)$ and $\mu_2 = (n_2P_2 + a_2)/(n_2 + a_2 + b_2)$. Evaluated under the null and alternative hypotheses, distributions **(3.1)** and **(3.2)** are obtained.

Notably, these two distributions are very similar to distributions of the frequentist test statistic, T_F, under H_0 and H_1. Thus, relationships among Bayesian statistic T_B, decision threshold c_B and type I and II error rates α_B and β_B can be established. Now, error probabilities, α_B and β_B can be represented in terms of the null and the alternative distributions of η and c_B.

$$\begin{aligned} \alpha_B &= 1 - F_{\eta,H_0}\left(\Phi^{-1}(c_B)\right), \\ \beta_B &= F_{\eta,H_1}\left(\Phi^{-1}(c_B)\right), \end{aligned} \tag{3.3}$$

where $F_\eta(\cdot)$ denotes the standard normal CDF function. Note that, unlike traditional hypothesis distribution, the distribution of η under the null and alternative represent different normal distributions, but applying transformations $\left(\eta - \mu_\eta\right)/\sigma_\eta$ results in an identical standard normal variable. Thus, the following equation can be derived from **(3.3)**:

$$\Phi^{-1}(1-\alpha_B)\sigma_{\eta,H_0} + \mu_{\eta,H_0} = \Phi^{-1}(\beta_B)\sigma_{\eta,H_1} + \mu_{\eta,H_1}. \tag{3.4}$$

By setting $\alpha_B = \alpha_F$ and $\beta_B = \beta_F$ and replace n_1 with $r * n_2$ in equation **(3.4)**, it is possible to solve the sample size for the control arm. The calculation of **(3.4)** can be solved using an iterative root-finding process. The threshold value c_B can be then calculated using either of the equations in **(3.3)** once the sample size is defined. That is:

$$c_B = \Phi\left(F^{-1}_{\eta,H_0}(1-\alpha)\right) = \Phi\left(F^{-1}_{\eta,H_1}(\beta)\right). \tag{3.5}$$

The resulting sample sizes n_1 and n_2 allow us to generate the desired type I and II error rates.

Sometimes the planned optimal sample size is unachievable because of logistic constraints or too few available patients in clinical trials. In this situation, the maximum sample size and the type I error rate are often pre-defined, and we determine the power of the study through power analysis. In Bayesian methods, the Bayesian power of the trial can be calculated as follow:

$$\varphi(n_B) = 1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha_B)\sigma_{\eta,H_0} + \mu_{\eta,H_0} - \mu_{\eta,H_1}}{\sigma_{\eta,H_1}}\right). \tag{3.6}$$

3.2 Results

Table 3.1 provides an example in which the Bayesian sample size and threshold provides values which are equivalent to desired frequentist type I and II error rates. Assuming that $P_1 = 0.65,\ P_2 = 0.5,\ r = 1,\ \Delta = 0,$, with $\alpha = 0.025$ or 0.05 and $\beta = 0.1, 0.15$ or 0.2 in this example. Note that under the null hypothesis $P_1 = P_2 = 0.5$. **Table 3.1** show that generated Bayesian type I and II error rates are consistently matched with the desired values for different values of α, β and Bayesian prior distributions. The actual values may slightly deviate from the desired values due to rounding up in the root solution to the equation **(3.4)** when calculating the sample size. When different priors are used, the null distribution **(3.1)** and the alternative distribution **(3.2)** may shift as to satisfy the error probabilities. Accordingly, the threshold c_B changes. All values of c_B are large numbers and are close to 1 in **Table 3.1**, which is similar to what Zaslavsky (2012) suggested using in a Bayesian hypothesis test.

Also, if n_B and c_B are already known, we can obtain α_B and β_B through the relationships defined in equation **(3.3)**. That means if a Bayesian design is presented, a parallel frequentist design can be found to produce the same error rates. To derive the frequentist design is relatively simple when comparing to the Bayesian design. Once α_F and β_F are known, the sample size calculation formula is straightforward and has been demonstrated by Lachin(Lachin, 1981). Frequentist

threshold, c_F, is defined as $Z_{1-\alpha_F}$. **Table 3.2** illustrate an example of finding equivalent frequentist designs for a given Bayesian design. Still, it is assumed that $P_1 = 0.65, P_2 = 0.5, r = 1, \Delta = 0$. Furthermore, we evaluate the parallel between frequentist and Bayesian methods for Bayesian total sample size of 250 or 500 and $c_B = 0.95$. Finally, corresponding n_F and c_F for the frequentist design are obtained. The resulting type I and II error rates are still the same between frequentist and Bayesian methods. Notably, frequentist sample sizes are different from Bayesian sample sizes when an informative prior distribution is used.

Figure 3.1 and **3.2** show comparisons of power between frequentist and Bayesian methods and Bayesian type I error rates, with the total sample size ranges from 150 to 300. Assuming that $P_1 = 0.65$, $P_2 = 0.5$, $r = 1$, $\Delta = 0$ and $\alpha = 0.05$. Further, we evaluate Bayesian powers under three prior distribution combinations: $Beta(1,1)$ for both experimental and control arms, $Beta(14, 6)$ for the experimental arm and $Beta(10, 10)$ for the control arm, and $Beta(6, 14)$ for the experimental arm and $Beta(10, 10)$ for the control arm. The last two indicate an enthusiastic and a pessimistic opinion about the treatment effect, respectively. The corresponding Bayesian threshold c_B is determined using **(2.5)**, given the sample size.

As shown in **Figure 3.1**, the Bayesian power is close to the frequentist power when applying non-informative priors to both treatment arms. However, the Bayesian power is not equivalent to the frequentist power when using informative prior distributions. The Bayesian power is higher with an enthusiastic prior than that of a pessimistic prior. Nevertheless, the type I error rate remains almost the same between the frequentist and the Bayesian methods as shown in **Figure 3.2**. Hence, our approach unifies only the type I error rate when frequentist and Bayesian sample sizes are set to the same value. In addition, this power calculation formula is useful in a sensitivity analysis. When observing treatment effect estimates $\widehat{P_1}$ and $\widehat{P_2}$ which deviate from

assumed treatment effects P_1 and P_2, we can calculate the power by replacing $P_1 = \widehat{P_1}$ and $P_2 = \widehat{P_2}$ in distributions of η under H_0 and H_1.

Frequentist results and Bayesian results of our unified approach are compared through a numerical study. In the numerical study, several scenarios are investigated including a small treatment effect $P_1 = 0.65$ and $P_2 = 0.5$ versus large treatment effect $P_1 = 0.7$ and $P_2 = 0.4$ and equal allocation $r = 1$ versus unequal allocation $r = 2$ using different priors. Assuming that $\Delta = 0, \alpha = 0.05$ and $\beta = 0.15$, we consider non-informative, optimistic and pessimistic priors for the experimental treatment arm in the numerical study. The optimistic prior assumes the experimental treatment is effective. For the small treatment effect scenario, the optimistic prior for the experimental arm is set as $a_1/(a_1 + b_1) = 0.65$ while for the large treatment effect it is set as $a_1/(a_1 + b_1) = 0.8$. While the pessimistic prior treats a treatment effect is unlikely to be seen, and is set as $a_1/(a_1 + b_1) = 0.35$ for the small treatment effect and $a_1/(a_1 + b_1) = 0.2$ for the large treatment effect. The prior for the control arm is always centered at 0.5, that $a_2/(a_2 + b_2) = 0.5$. In addition, the strength of the prior is taken into account. The strength is represented as the ratio of the sum of prior parameters to the final sample size. A stronger prior has more influence on the posterior than a weaker prior. For the small treatment effect scenario, a strong prior is set as $a_i + b_i = 20$ and a weak prior is set as $a_i + b_i = 10$. While for the large treatment effect scenario, a strong prior is set as $a_i + b_i = 40, i = 1,2$ and a weak prior is set as $a_i + b_i = 10$. The results are shown in **Table 3.3**. The type I error rate and the power are generated from 10000 simulations using R for corresponding frequentist and Bayesian methods.

The resulting type I and II error rates are comparable between the frequentist and the Bayesian methods. In all scenarios, type I and II error rates are approximately the same for frequentist and Bayesian methods. Frequentist and the Bayesian sample sizes are approximately

equal when using non-informative priors for Bayesian methods. However, Bayesian sample sizes are different from frequentist sample sizes when informative priors are applied. When applying optimistic prior distributions, Bayesian methods require smaller sample sizes than frequentist methods regardless of whether an unequal allocation ratio is used. For pessimistic priors, Bayesian methods require larger sample sizes than frequentist methods in equal allocation scenarios. In contrast, Bayesian sample sizes are smaller than frequentist sample sizes in unequal allocation cases.

3.3 Conclusions

In this paper, a novel unified approach for frequentist and Bayesian one-sided hypothesis tests in two-arm fixed-sample trials with binary endpoints was developed. It is assumed that the Bayesian prior follows a conjugated beta distribution. The approach unified the approaches controlling for both type I and II error rates resulting from frequentist and Bayesian approaches by theoretically establishing a one-to-one mapping between these two methods. Unlike previous approaches that only controlled type I error rate in Bayesian clinical trials, the proposed unified approach preserves both type I and type II error rates at desired levels by incorporating Bayesian prior distributions into sample size and threshold determinations. Therefore, for Bayesian hypothesis testing, it is unnecessary to adjust the prior distribution to achieve the unification with frequentist methods. In addition, the assumptions are relaxed on priors and hypotheses to make the approach more generalizable for trials with binary endpoints. Conjugate beta prior parameters need not necessarily be integers and hypotheses include a parameter to represent the minimal clinically important difference. Also, theoretical methods for determining the sample size, threshold and power are provided. These methods do not require intensive computations or simulations. Under normal approximation to the posterior probability distribution, sample size and threshold values

were explicitly obtained. As mentioned in **Section 3.2**, our approach could be extended to any distribution. The normal approximation was chosen because in most clinical trials this seems reasonable. (To examine the validity of normal approximation, especially under small samples, a small simulation study was performed (See **Figure 3.3**). The approximation seems reasonable.)

3.4 Appendix

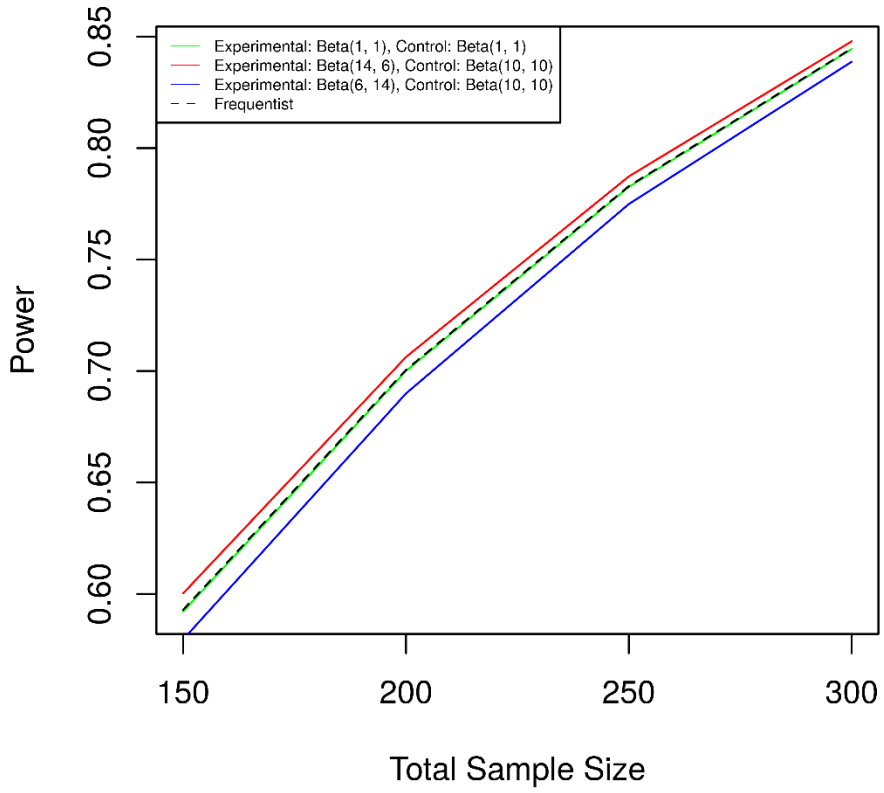


Figure 3.1 Frequentist and Bayesian power comparisons with total sample sizes range from 150 to 300. We assume that $P_1 = 0.65$, $P_2 = 0.5$, $r = 1$, $\Delta = 0$ and $\alpha = 0.05$. For Bayesian methods, we evaluate their powers under three prior distribution combinations: $Beta(1, 1)$ for both arms (the green curve), $Beta(14, 6)$ for the experimental arm and $Beta(10, 10)$ for the control arm (the red curve), and $Beta(6, 14)$ for the experimental arm and $Beta(10, 10)$ for the control arm (the blue curve). The black dotted curve refers to frequentist powers.

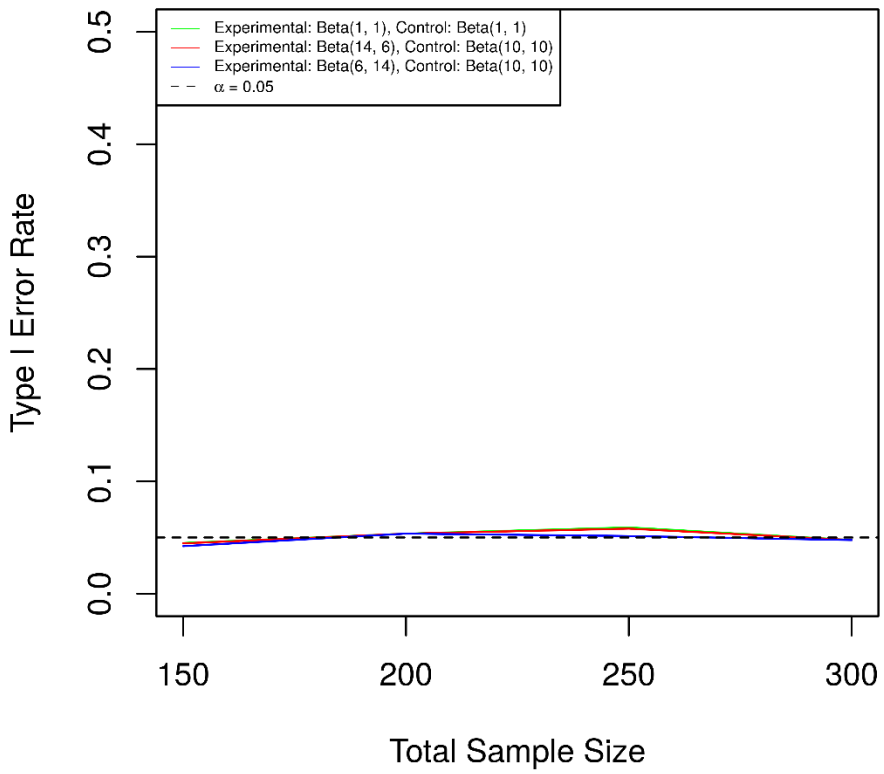


Figure 3.2 Bayesian type I error rates under the same assumptions as **Figure 3.1**.

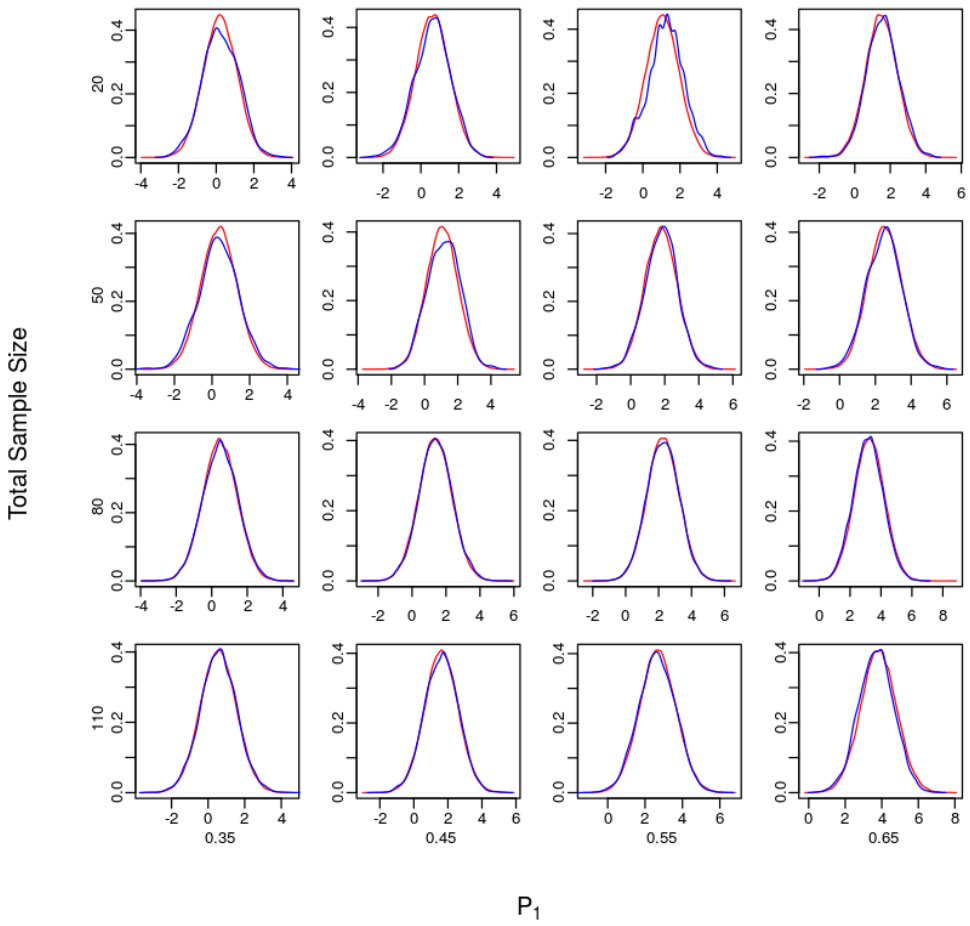


Figure 3.3 Evaluation of normal approximation under different sample sizes and treatment effect sizes. Assuming that $P_2 = 0.3$, $r = 1$, $\Delta = 0$ and non-informative prior distributions for both treatment arms. The total sample size varies from 20 to 110 and P_1 varies from 0.35 to 0.65. The blue line represents the density curve of transformed posterior probabilities under the alternative hypothesis, η_{H_1}, and the red line represents the normal distribution used as the approximation to the distribution of η_{H_1}. From the figure below, it can be seen that the distribution of η can be well approximated when the sample size is greater than 20.

Table 3.1 Some scenarios to evaluate the Bayesian type I and II error probabilities α_B and β_B under different priors. Assuming that $P_1 = 0.65,\ P_2 = 0.5,\ r = 1,\ \Delta = 0$ with $\alpha = 0.025$ or 0.05, $\beta = 0.1, 0.15$ or 0.2, we report power here but the type II error rate can be calculated as 1 – power.

Treatment arm 1 prior	Treatment arm 2 prior	Bayesian sample size	Bayesian decision threshold (c_B)	Desired type I error rate	Generated type I error rate	Desired power	Generated power
$Beta(1,1)$	$Beta(1,1)$	264	0.949	0.05	0.0477	0.8	0.801
$Beta(7,3)$	$Beta(6,4)$	306	0.956	0.05	0.0482	0.85	0.843
$Beta(3,7)$	$Beta(5,5)$	368	0.919	0.05	0.0502	0.9	0.911
$Beta(3,7)$	$Beta(5,5)$	338	0.955	0.025	0.0245	0.8	0.813
$Beta(1,1)$	$Beta(1,1)$	382	0.974	0.025	0.0233	0.85	0.844
$Beta(7,3)$	$Beta(6,4)$	446	0.978	0.025	0.0253	0.9	0.908

Table 3.2 Some scenarios to evaluate the frequentist sample size under different priors, Bayesian sample size and threshold. Assuming that $P_1 = 0.65, P_2 = 0.5, r = 1, \Delta = 0$, we set a total sample size for the Bayesian method to be 250 and 500, with $c_B = 0.95$. The frequentist type I error rate and power are simulated using the calculated frequentist sample size.

Treatment arm 1 prior	Treatment arm 2 prior	Bayesian sample size	Frequentist sample size	Bayesian type I error rate	Frequentist type I error rate	Bayesian power	Frequentist power
$Beta(1,1)$	$Beta(1,1)$	250	250	0.0493	0.0501	0.782	0.784
$Beta(7,3)$	$Beta(6,4)$	250	252	0.0574	0.0562	0.804	0.813
$Beta(3,7)$	$Beta(5,5)$	250	246	0.0253	0.0253	0.674	0.681
$Beta(1,1)$	$Beta(1,1)$	500	500	0.0496	0.0498	0.963	0.967
$Beta(7,3)$	$Beta(6,4)$	500	502	0.0565	0.0564	0.968	0.969
$Beta(3,7)$	$Beta(5,5)$	500	496	0.0320	0.0325	0.943	0.946

Table 3.3a Results of the numerical study comparing frequentist and Bayesian methods for the large treatment effect ($P_1 = 0.65\ and\ P_2 = 0.5$). Assuming that $\Delta = 0$ with $\alpha = 0.05$ and $\beta = 0.15$, we report the sample size of the control group (treatment group 2). Let "F" denote frequentist methods and "B" denote Bayesian methods.

Allocation ratio	Treatment arm 1 prior	Treatment arm 2 prior	Sample size of the control arm		Threshold		Type I error rate		Power	
			F	B	F	B	F	B	F	B
1:1	$Beta(1,1)$	$Beta(1,1)$	153	153	1.64	0.949	0.0503	0.0517	0.848	0.843
	$Beta(6.5,3.5)$	$Beta(5,5)$		153		0.961		0.05		0.842
	$Beta(3.5,6.5)$	$Beta(5,5)$		154		0.924		0.048		0.841
	$Beta(26,14)$	$Beta(20,20)$		152		0.981		0.0504		0.841
	$Beta(14,26)$	$Beta(20,20)$		157		0.807		0.0509		0.856
2:1	$Beta(1,1)$	$Beta(1,1)$	114	115		0.949	0.0496	0.05	0.85	0.846
	$Beta(6.5,3.5)$	$Beta(5,5)$		110		0.956		0.0485		0.856
	$Beta(3.5,6.5)$	$Beta(5,5)$		111		0.93		0.0514		0.846
	$Beta(26,14)$	$Beta(20,20)$		94		0.972		0.0523		0.855
	$Beta(14,26)$	$Beta(20,20)$		97		0.83		0.0506		0.861

Table 3.3b Results of the numerical study comparing frequentist and Bayesian methods for the small treatment effect ($P_1 = 0.7$ and $P_2 = 0.4$). Assuming that other parameters equal to those in **Table 3.3a**.

Allocation ratio	Treatment arm 1 prior	Treatment arm 2 prior	Sample size of the control arm		Threshold		Type I error rate		Power	
			F	B	F	B	F	B	F	B
1:1	$Beta(1,1)$	$Beta(1,1)$	36	37	1.64	0.947	0.0505	0.0544	0.851	0.834
	$Beta(8,2)$	$Beta(5,5)$		35		0.981		0.0496		0.852
	$Beta(2,8)$	$Beta(5,5)$		38		0.805		0.0492		0.848
	$Beta(16,4)$	$Beta(10,10)$		33		0.993		0.0492		0.837
	$Beta(4,16)$	$Beta(10,10)$		38		0.582		0.0414		0.839
2:1	$Beta(1,1)$	$Beta(1,1)$	28	27		0.944	0.0521	0.0477	0.846	0.844
	$Beta(8,2)$	$Beta(5,5)$		22		0.963		0.0511		0.847
	$Beta(2,8)$	$Beta(5,5)$		23		0.794		0.0543		0.848
	$Beta(16,4)$	$Beta(10,10)$		18		0.98		0.0458		0.848
	$Beta(4,16)$	$Beta(10,10)$		20		0.534		0.051		0.835

CHAPTER FOUR: SPECIFIC AIM 2

4.1 Introduction

Following the unified approach for frequentist and Bayesian methods in fixed-sample clinical trials in **Chapter 3**, we propose another novel and theoretical-based unified approach for group-sequential clinical trials in this chapter. Similarly, the unified approach aimed to achieve the same type I and II error rates between frequentist and Bayesian methods for all analyses in group sequential trials. These analyses include interim analyses and the final analysis. When a frequentist design is given, the unified approach determines the Bayesian stopping boundaries and sample size through a theoretical approach. When a Bayesian design is given, the unified approach calculates the frequentist type I and II error rates, leading to a corresponding frequentist design. Alpha spending functions for controlling the overall type I error rate and beta spending function for controlling the overall type II error rate are used for the unified approach in **Aim 2**. Similar to **Aim 1**, it is assumed that the group sequential trial still has two arms and binary outcomes. Beta conjugate priors and decision making based on posterior probabilities for treatment difference are used in Bayesian group sequential methods are used as well. Based on the approximate distributions **(3.1)** and **(3.2)**, a joint distribution of Bayesian posterior probabilities is derived. **Section 4.2** explained the novel unified approach for frequentist and Bayesian group sequential methods. In Section **4.3**, numerical results that show the unified approach achieves the goal under various scenarios are present. The method is also applied to a trial application.

4.2 Methods

Assume the same parameterizations as in **Chapter 3** and consider a group sequential trial comparing an experimental treatment arm to a control treatment arm that has J stages, that $J \geq 2$. Let subscript j denote the stage, $j = 1, \dots, J$, where an analysis is carried out at the end of each

stage. Let n_j denote the accumulated number of patients at the end of the *j*th stage. The one-sided hypotheses $H_0: \pi_1 - \pi_2 \leq \Delta$ against $H_1: \pi_1 - \pi_2 > \Delta$ are tested at the end of each stage.

Further, let $c_{E,j}$ and $c_{F,j}$ denote the efficacy and the futility boundary for the *j*th stage, respectively. The trial with both efficacy and futility boundaries shall continue if the test statistic T_j at *j*th analysis, satisfying $c_{F,j} < T_j < c_{E,j}, j = 1, \dots, J-1$. Otherwise, the trial will be terminated early because the pre-specified stopping boundary has been crossed. The trial is stopped for overwhelming benefits of the experimental treatment if $T_j > c_{E,j}$, or it is stopped for sufficient negative treatment effect if $T_j < c_{F,j}$. The relationships among stopping boundaries, type I and II error rates have been illustrated in **Section 1.2.3**.

Also, the expected sample size formula has been provided in **Section 1.2.5**, that is:

$$\text{ESS} = \sum_{j=1}^{J-1} n_j Pr\left(T_j < c_{F,j} \cup T_j > c_{E,j}\right) + n_J Pr\left(\bigcap_{j}^{j-1} c_{F,j} \leq T_j \leq c_{E,j}\right).$$

Alpha and beta spending functions are considered for the unified approach. The alpha spent at each interim analysis is a function of the information fraction, denoted as τ. The information fraction of the *j*th stage, τ_j, is defined as n_j/n_J, where n_J is the maximum sample size required in the clinical trial. Further, three alpha spending functions introduced in **Section 1.2.4** are considered here:

(1). $\alpha_1(\tau) = 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{\tau}\right)$, O'Brien-Fleming

(2). $\alpha_2(\tau) = \alpha ln(1 + (e-1)\tau)$, Pocock

(3). $\alpha_3(\tau) = \alpha\tau^{\theta}$, for $\theta > 0$, Power

α_1 and the increments $\alpha_2 - \alpha_1, \dots, \alpha_j - \alpha_{j-1}$ are the type I error rate allowed for the analysis at stage $1,2, \dots, J$.Similarly, the above functions can be applied to spend beta by replacing the α parameter with β parameter.

Same as **Aim 1**, conjugate prior distributions $Beta(a_i, b_i), i = 1,2$, for the experimental arm and the control arm is considered. However, the posterior distribution is updated at every stage. At the end of the *j*th stage, π_{ij} follows a beta posterior distribution $Beta\left(a_i + y_{ij}, b_i + n_{ij} - y_{ij}\right)$, where n_{ij} and y_{ij} are the number of patients and the number of patients with successful outcomes for the *i*th arm by the *j*th stage, and $n_j = \sum_i^2 n_{ij}$. The test statistic, T_j at the *j*th analysis, is the posterior probability that the experimental arm is superior to the control arm, $T_j = P_j\left(\pi_{1j} - \pi_{2j} > \Delta\right)$. Similar to classical group sequential procedures, the trial will be terminated for sufficient evidences of efficacy when $P_j\left(\pi_{1j} - \pi_{2j} > \Delta\right)$ is greater than the c_{Ej}, and will be terminated for futility when the posterior probability is smaller than c_{Fj}. Though type I and II error rates are not naturally defined in Bayesian methodologies, Monte Carlo simulations can be used to calculate type I and II error rates. For instance, to calculate the type I error rate for a Bayesian group sequential trials with double boundaries at the *j*th stage, *L* samples are drawn from the Bayesian posterior distributions under the null distribution, and then calculate the proportion of samples that the trial is stopped for efficacy:

$$\alpha_j = \sum_{l=1}^{L} I_{H_0}\left(\bigcap_{j^*}^{j-1} c_{F,j^*} \le P_{j^*,l}\left(\pi_{1j^*,l} - \pi_{2j^*,l} > \Delta\right) \le c_{E,j^*} \cap P_{j,l}\left(\pi_{1j,l} - \pi_{2j,l} > \Delta\right) > c_{E,j}\right)/L,$$

where $I(.)$ is an indicator function and the overall type I error rate is the summation of $\alpha_1, \alpha_2, \dots, \alpha_J$. Similarly, the type II error rate estimated from Monte Carlo simulations is defined as follows:

$$\beta_j = \sum_{l=1}^{L} I_{H_1}\left(\bigcap_{j^*}^{j-1} c_{F,j^*} \le P_{j^*,l}\left(\pi_{1j^*,l} - \pi_{2j^*,l} > \Delta\right) \le c_{E,j^*} \cap P_{j,l}\left(\pi_{1j,l} - \pi_{2j,l} > \Delta\right) < c_{F,j}\right)/L,$$

and the overall type II error rate is the summation of $\beta_1, \beta_2, \ldots, \beta_J$. Nevertheless, those simulation-based calculations cost considerable amounts of time, especially at the designing stage when clinical trialists want to evaluate a number of Bayesian stopping boundaries.

Therefore, a more efficient way is to derive a theoretical method to calculate the type I and II error rates directly from the posterior distributions of $T_1, \ldots, T_J$. Notably, the posterior distributions of $T_1, \ldots, T_J$ are highly skewed and have no closed-form solutions. However, in **Chapter 3** we have showed that the standard normal quantile transformation of the posterior probability $T_\cdot$, denoted as $\eta_\cdot$, can be well approximated by a normal distribution in two-arm fixed sample trials with binary endpoints, when the sample size is sufficient large. As such, in group-sequential methods, the distribution of $\eta_j = \Phi(T_j)$ can also be approximated by a normal distribution, that is,

$$\eta_j \sim N\left(\frac{\mu_{1j} - \mu_{2j} - \Delta}{\sqrt{\frac{\mu_{1j}(1-\mu_{1j})}{n_{1j} + a_1 + b_1 + 1} + \frac{\mu_{2j}(1-\mu_{2j})}{n_{2j} + a_2 + b_2 + 1}}}, \frac{\frac{n_{1j}P_2(1-P_2)}{(n_{1j} + a_1 + b_1)^2} + \frac{n_{2j}P_2(1-P_2)}{(n_{2j} + a_2 + b_2)^2}}{\frac{\mu_{1j}(1-\mu_{1j})}{n_{1j} + a_1 + b_1 + 1} + \frac{\mu_{2j}(1-\mu_{2j})}{n_{2j} + a_2 + b_2 + 1}}\right), \quad (4.1)$$

where $\mu_{1j} = (n_{1j}P_1 + a_1)/(n_{1j} + a_1 + b_1)$ and $\mu_{2j} = (n_{2j}P_2 + a_2)/(n_{2j} + a_2 + b_2)$. Similar to frequentist group sequential clinical trials(Jennison & Turnbull, 2000), the normal approximation works well because sample sizes are enough large. Note, $(\eta_1, \eta_2, \ldots, \eta_J)$ follows a multivariate normal distribution(Jennison & Turnbull, 2000). The covariance between the transformed variables at any two stages j_1 and j_2, $1 \le j_1 \le j_2 \le J$, is

$$Cov(\eta_{j1}, \eta_{j2}) = \sqrt{\frac{\sigma_{j_1}^2 \sigma_{j_2}^2 \sigma_{Fj_2}^2}{\sigma_{Fj_1}^2}}, 1 \le j_1 \le j_2 \le J,$$

where $\sigma_{Fj_1}^2 = n_{1j1}P_1(1-P_1)/{n_{1j1}}^2 + n_{2j1}P_2(1-P_2)/{n_{2j1}}^2$ and $\sigma_{Fj_2}^2 = n_{1j2}P_1(1-P_1)/{n_{1j2}}^2 + n_{2j2}P_2(1-P_2)/{n_{2j2}}^2$. A proof for the covariance is shown below.

Proof 4.1

Suppose there is a frequentist group sequential design with the same P_i, α, β and n_{ij}, $i = 1,2; j = 1, \dots, J$. Let $Z_{j1}\ and\ Z_{j2}$ denote the standard test statistics at the $j_1 th$ and the $j_2 th$ stages, $1 \le j_1 < j_2 \le J$. The covariance between $Z_{j1}\ and\ Z_{j2}$ has been derived by Jennison and Turnbull (2001), that:

$$Cov(Z_{j1}, Z_{j2}) = \sqrt{\sigma_{Fj_2}^2/\sigma_{Fj_1}^2}.$$

Further, write covariance of η_{j_1} and η_{j_2} as the covariance of functions of Z_{j1} and Z_{j2}:

$$Cov(\eta_{j_1}, \eta_{j_2}) = Cov\left(Z_{j_1}\sigma_{j_1} + \mu_{\eta_{j_1}}, Z_{j_2}\sigma_{j_2} + \mu_{\eta_{j_2}}\right) = \sigma_{j_1}\sigma_{j_2}Cov(Z_{j1}, Z_{j2}).$$

Replace $Cov(Z_{j1}, Z_{j2})$ with $\sqrt{\sigma_{Fj_2}^2/\sigma_{Fj_1}^2}$, the covariance of η_{j_1} and η_{j_2} is obtained as:

$$Cov(\eta_{j1}, \eta_{j2}) = \sqrt{\frac{\sigma_{j_1}^2 \sigma_{j_2}^2 \sigma_{Fj_2}^2}{\sigma_{Fj_1}^2}}.$$

The resulting multivariate normal distribution allows us to calculate Bayesian type I and II error rates using a numerical approach. Bayesian type I and II error rates at the *j*th stage are defined as:

$$\alpha_j = Pr_{H_0}\left(\bigcap_{j^*}^{j-1} \Phi\left(c_{F,j^*}\right) \le \eta_{j^*} \le \Phi\left(c_{E,j^*}\right) \cap \eta_j > \Phi\left(c_{E,j}\right)\right), \tag{4.2}$$

$$\beta_j = Pr_{H_1}\left(\bigcap_{j^*}^{j-1} \Phi\left(c_{F,j^*}\right) \le \eta_{j^*} \le \Phi\left(c_{E,j^*}\right) \cap \eta_j < \Phi\left(c_{F,j}\right)\right). \tag{4.3}$$

With the formulae to calculate Bayesian type I and II error rates, it is possible to obtain the same type I and II error rates for each analysis for frequentist and Bayesian group sequential methods. Let subscript F denote parameters for frequentist methods and subscript B denote parameters for Bayesian methods. Unification between frequentist and Bayesian methods is achieved by solving the following equations for parameters of one method when parameters of the other are given:

$$\alpha_{F,1} = \alpha_{B,1} \text{ and } \alpha_{F,j+1} - \alpha_{F,j} = \alpha_{B,j+1} - \alpha_{B,j}, j = 1, \dots, J-1. \tag{4.4}$$

$$\beta_{F,1} = \beta_{B,1}, \text{ and } \beta_{F,j+1} - \beta_{F,j} = \beta_{B,j+1} - \beta_{B,j}, j = 1, \dots, J-1. \tag{4.5}$$

$\alpha_{F,1}$ and $\beta_{F,1}$ are type I and II error rates distributed at the *j*th analysis using spending functions. Note, there are several other methods than spending functions to define stopping boundaries, e.g. boundary computation for triangular test(Whitehead & Stratton, 1983) and the recursive numerical algorithm proposed by Armitage et al.(P Armitage et al., 1969) But it is beyond the scope of this paper and will not be discussed in detail. But the same unification results can always be achieved between frequentist and Bayesian group sequential methods as long as equations **(4.4)** and **(4.5)** are satisfied. Notably, the stopping probability also becomes the same for frequentist and Bayesian methods for each analysis if **(4.4)** and **(4.5)** are both met.

If a Bayesian group sequential design is given, corresponding frequentist group sequential methods to achieve the same type I and II error rates are derived using the unified approach. For

instance, in a Bayesian group sequential trial with double boundaries, $\alpha_{B,j}$ and $\beta_{B,j}, j = 1, \dots, J$, can be calculated using equations **(4.2)** and **(4.3)**. Frequentist stopping boundaries and sample size are derived for given $\alpha_{F,1}, \dots, \alpha_{F,J}$ and $\beta_{F,1}, \dots, \beta_{F,j}$, that satisfy equations **(4.4)** and **(4.5)**. Jennison and Turnbull(Jennison & Turnbull, 2000) demonstrated how to determine stopping boundaries and the maximum sample size for frequentist group sequential methods.

Likewise, corresponding Bayesian group sequential methods are derived when a frequentist group sequential design is given. Similar to frequentist group sequential methods, Bayesian stopping boundaries are computed sequentially. However, detailed steps to determine Bayesian stopping boundaries considering prior influences may differ from frequentist steps. In equation **(4.1)**, the mean of η_j under the null contains the sample size parameter and is not centered at zero if informative priors are applied. Thus, the Bayesian sample size is estimated simultaneously when calculating Bayesian stopping boundaries. Algorithms for a single Bayesian boundary and double Bayesian boundaries computations are illustrated as follows.

In group sequential methods with a single boundary, only one of efficacy and futility boundaries are specified, while the other boundaries are not defined. Resulting trials are stopped either for efficacy or futility. Use efficacy boundaries as an example, alpha spending functions are applied to control the overall type I error of group sequential trials. The relationship among the alpha spent, the efficacy boundary and the Bayesian sample size at the first stage is established below:

$$Pr_{\mathrm{H}_0}\left(\eta_1 \geq \Phi\left(c_{E,1}\right)\right) = \int_{c_{E,1}}^{\infty} f(\eta_1) d\eta_1 = \alpha(\tau_1), \tag{4.6}$$

where $f(\eta_1)$ is the marginal probability density function of η_1. For the following stages, the increment of the alpha spent between jth and $(j+1)$th stages should equal to the integral of the multivariate normal distribution of $\eta_1, \dots \eta_{j+1}$.

$$Pr_{H_0}\left(\eta_1 < \Phi(c_{E,1}), \dots, \eta_{j-1} < \Phi(c_{E,j-1}), \eta_j \geq \Phi(c_{E,j})\right) = \tag{4.7}$$

$$\int_{-\infty}^{c_{E,1}} \dots \int_{c_{E,j}}^{\infty} f(\eta_1, .., \eta_j) d\eta_1 \dots d\eta_j = \alpha(\tau_j) - \alpha(\tau_{j-1}),\ j = 2, \dots, J.$$

Bayesian power is defined as:

$$1 - \beta = Pr_{\mathrm{H}_1}\left(\bigcup_j^J \eta_j > \Phi(c_{E,j})\right). \tag{4.8}$$

The algorithm to compute the efficacy boundaries and Bayesian sample size works as follows:

1. Calculate the type I error rate spent at stage *j* in the Bayesian group sequential method, that is, $\alpha_B(\tau_1), \alpha_B(\tau_2) - \alpha_B(\tau_1), \dots, \alpha_B(\tau_J) - \alpha_B(\tau_{J-1})$. In the unified approach, the type I error rate spent at each stage in the Bayesian method is set to the same as that in the frequentist method, as stated in equation (4).
2. Select a possible range of maximum sample size from n^0 to n^M, $n^0 \leq n^M$, denoted as $\{n^0, n^1, \dots, n^M\}$. Choose an initial value of the maximum sample size, e.g. $n = n^0$.
3. Compute the efficacy boundaries for the first interim analysis, $c_{E,1}$, by plugging sample size parameter n into equation **(4.6)**. Solve $c_{E,2}$,…, $c_{E,\mathrm{J}}$ sequentially using equation **(4.7)**.
4. Compute the power $1 - \beta^*$, using the equation **(4.8)**.
5. Since the power is a monotone function of the sample size when holding the type I error rate fixed. Iterate over values in $\{n^0, n^1, \dots, n^M\}$, repeat steps 3 and 4 until the difference

between $1-\beta^*$ and to $1-\beta$ is smaller than some pre-specified margin of error, ε, that is, $|\beta^*-\beta|<\varepsilon$. ε is a small number, i.e. $1e-4$. The final *n* is the maximum sample size and $c_{E,1}$, $c_{E,2}$,…, $c_{E,J}$ are corresponding efficacy boundaries.

Optimizations such as bisection methods can be implemented to reduce the time to search for the maximum sample size parameter *n*.

The bisection method proceeds as follows:

1) For the possible range of maximum sample size, $\{n^0, n^1, \ldots, n^M\}$, define the minimum value as *MIN* = n^0 and the maximum value as *MAX* = n^M. Define the midpoint as the mean of *MIN* and *MAX* rounded down to the nearest integer, that is, *MID* = floor$\left({}^1/_2\,(n^0+n^M)\right)$. Compute Bayesian powers $1-\beta^*$ in step 4 for *MAX* and *MIN*, respectively. Calculate differences between resulting Bayesian powers and the desired power $1-\beta$ for *MAX* and *MIN*, denoted as $D(MAX)$ and $D(MIN)$. Note, $D(MAX)$ and $D(MIN)$ should have opposite signs.
2) Compute the difference between Bayesian power calculated in step 4 using MID and the desired power, denoted as $D(MID)$.
3) If $D(MID)$ has the opposite sign of $D(MAX)$, update the *MIN* as $MIN = MID+1$. Otherwise, update the *MAX* as $MAX = MID-1$. Next, the new *MID* value is updated as floor$\left({}^1/_2\,(n^0+n^M)\right)$.
4) Repeat step 2 and 3 until $MAX \le MIN$. The final maximum sample size is *n* = *MIN*.

Steps of calculating for futility boundaries are very much the same as for efficacy boundaries. The few changes in steps for futility boundaries computation are to replace the alpha spending function

with the beta spending function and search for the maximum sample size until the overall type I error rate $Pr_{\mathrm{H}_0}\left(\bigcap_j^{J-1}\eta_j \geq \Phi(c_{F,j}) \cap \eta_J \geq \Phi(c_{F,J})\right)$ converges to the desired type I error rate.

2.5. Bayesian Group Sequential Methods with Double Boundaries

Many group sequential trials implement both efficacy and futility boundaries. Alpha and beta spending functions are applied to maintain the overall type I and II error rates for trials. Stopping boundary values for the first interim analysis are computed using the following equations:

$$Pr_{H_0}\left(\eta_1 \geq \Phi(c_{E,1})\right) = \int_{c_{E,1}}^{\infty} f(\eta_1)d\eta_1 = \alpha(\tau_1), \tag{4.9}$$

$$Pr_{H_1}\left(\eta_1 < \Phi(c_{F,1})\right) = \int_{-\infty}^{c_{F,1}} f(\eta_1)d\eta_1 = \beta(\tau_1). \tag{4.10}$$

Notably, there are two different ways to calculate the following boundary values. The first approach binds efficacy boundaries with futility boundaries. That is, the trial is terminated when any futility boundary has been crossed. Following equations are used to calculate efficacy and futility boundaries for the *j*th stage, $j = 2, \ldots, J$.

$$\begin{aligned} &Pr_{H_0}\left(\Phi(c_{F,1}) < \eta_1 < \Phi(c_{E,1}), \ldots, \Phi(c_{F,j-1}) < \eta_{j-1} < \Phi(c_{E,j-1}), \eta_j \right. \\ &\qquad \left. \geq \Phi(c_{E,j})\right) \\ &= \int_{c_{F,1}}^{c_{E,1}} \ldots \int_{c_{E,j}}^{\infty} f(\eta_1, .., \eta_j)d\eta_1 \ldots d\eta_j = \alpha(\tau_j) - \alpha(\tau_{j-1}), \end{aligned} \tag{4.11}$$

$$\begin{aligned} &Pr_{H_1}\left(\Phi(c_{F,1}) < \eta_1 < \Phi(c_{E,1}), \ldots, \Phi(c_{F,j-1}) < \eta_{j-1} < \Phi(c_{E,j-1}), \eta_j \right. \\ &\qquad \left. \geq \Phi(c_{E,j})\right) \end{aligned} \tag{4.12}$$

$$= \int_{c_{F,1}}^{c_{E,1}} \dots \int_{c_{E,j}}^{\infty} f(\eta_1, .., \eta_j) d\eta_1 \dots d\eta_j = \beta(\tau_j) - \beta(\tau_{j-1}).$$

And the power is defined as

$$1 - \beta = 1 - \sum_{j=1}^{J} Pr_{H_1}\left(\bigcap_{j^*}^{j-1} \Phi(c_{F,j^*}) \leq \eta_{j^*} \leq \Phi(c_{E,j^*}) \cap \eta_j < \Phi(c_{F,j}) \right). \quad (4.13)$$

Oppositely, the second approach considers unbinding futility boundaries. The trial can continue even if a futility boundary has been crossed. In this approach, steps to obtain efficacy boundaries are the same as equation **(4.7)** in the single efficacy boundary calculation. While futility boundaries computation is the same as equation **(4.12)**.

The algorithm to compute binding stopping boundaries and the maximum sample size in Bayesian group sequential methods work as follows:

1. Calculate the type I and II error rates spent at stage j, that is, $\alpha_B(\tau_1)$, $\alpha_B(\tau_2) - \alpha_B(\tau_1), \dots, \alpha_B(\tau_J) - \alpha_B(\tau_{J-1})$ and $\beta_B(\tau_1)$, $\beta_B(\tau_2) - \beta_B(\tau_1), \dots, \beta_B(\tau_J) - \beta_B(\tau_{J-1})$. In the unified approach, type I and II error rates spent at each stage in the Bayesian method is set to the same as that in the frequentist method, as stated in equation **(4.4)** and **(4.5)**.

2. Select a possible range of maximum sample size from n^0 to n^M, $n^0 \leq n^M$, denoted as $\{n^0, n^1, \dots, n^M\}$. Choose an initial value of the maximum sample size, e.g. $n = n^0$.
3. Compute efficacy and futility boundaries for the first interim analysis $c_{E,1}$ and $c_{F,1}$ by plugging sample size parameter n into equation **(4.9)** and **(4.10)**. Solve $c_{E,2}, c_{F,2} \dots, c_{E,J}$ sequentially using equation **(4.11)** and **(4.12)**, set $c_{F,J} = c_{E,J}$.
4. Compute the power using equation **(4.13)**.

5. Since the power is a monotone function of the sample size when holding the type I error rate fixed. Iterate over values in $\{n^0, n^1, \ldots, n^M\}$, repeat steps 3 and 4 until the difference between $1-\beta^*$ and to $1-\beta$ is smaller than some pre-specified margin of error, ε, that is, $|\beta^* - \beta| < \varepsilon$. ε is a small number, i.e. $1e-4$. The final n is the maximum sample size. $c_{E,1}$, $c_{E,2}$,…, $c_{E,J}$ and c_{F1}, $c_{F,2}$,…, $c_{F,J}$ are corresponding efficacy and futility boundaries.

For non-binding boundaries, step 3 is changed to compute efficacy boundaries $c_{E,1}$ and $c_{F,1}$ using equations **(4.7)** and **(4.12)**. Similarly, using the bisection algorithm can reduce the computation time to obtain the maximum sample size.

4.3 Results

First, the stopping boundaries are evaluated with different prior specifications. **Figure 4.1** shows efficacy boundaries for Bayesian group sequential trials assuming single boundaries. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$. O'Brien-Fleming boundary, Pocock boundary, and power boundaries with $\theta = 1, 2$ are compared under three different Bayesian prior combinations for treatment arms, including a non-informative prior, an optimistic prior and a pessimistic prior. $\mathrm{Beta}(1,1)$ distributions are applied to both treatment arms as non-informative priors. For the optimistic prior, it is considered $\mathrm{Beta}(13,7)$ for the experimental arm and $\mathrm{Beta}(10,10)$ for the control arm, so that the prior favors the experimental arm. For the pessimistic prior, $\mathrm{Beta}(5,15)$ and $\mathrm{Beta}(10,10)$ are assigned to the experimental treatment and control treatment, respectively. Thus, the prior assumes a treatment effect is unlikely to be observed. Consider three analyses are planned for the trial, two of which are interim analysis to be conducted when 1/3 and 1/2 patients are enrolled. The information fraction, τ, is used for drawing the x-axis in **Figure 4.1**. **Figure 4.2** show efficacy and futility stopping boundaries assuming binding double boundaries. The efficacy boundary and the futility boundary are either a

Pocock type of stopping boundaries, or an O'Brien-Fleming type of stopping boundaries. Pocock and O'Brien-Fleming boundaries are chosen because they are commonly used in group-sequential clinical trials. It is assumed that **Figure 4.2** uses the same P_1, P_2, r, Δ, α and β as in **Figure 4.1**.

When different priors are used, the null distribution and the alternative distribution may shift to satisfy the type I and type II probabilities. Therefore, Bayesian stopping boundary values vary for different Bayesian priors. In both figures, it can be seen that the efficacy boundary values are more rigorous for optimistic priors compared to those for pessimistic priors. Similar to frequentist stopping boundaries, O'Brien-Fleming and power with $\theta = 2$ efficacy boundaries are more stringent at early stages compared to Pocock and power with $\theta = 1$ boundaries in **Figure 4.1**. The changing patterns for four types of efficacy boundaries are similar for non-informative prior and optimistic prior in **Figure 4.1**. While it is interesting to find efficacy boundaries have a larger variation between stages for pessimistic prior than other two types of priors. Also, values of Pocock and power with $\theta = 1$ boundaries increases much from the first interim analysis to the second interim analysis. For double boundaries case in **Figure 4.2**, the futility boundaries for the optimistic prior is higher than that of the non-informative prior, and the futility boundaries for the pessimistic prior is lower than that of the non-informative prior. Similar to **Figure 4.1** when pessimistic prior is applied, stopping boundaries values vary much between stages and the Pocock efficacy boundaries increase from the first interim analysis to the second interim analysis. In addition, the trial is less likely to stop early using O'Brien-Fleming futility boundaries than Pocock futility boundaries, especially when pessimistic prior is used. Overall, both figures indicate Bayesian prior as well as the type of stopping boundaries should be carefully evaluated and chosen by clinical trialists, as the values and patterns of resulting stopping boundaries can be quite different.

Furthermore, Bayesian type I and II error rates and stopping probabilities are evaluated under the H_1 for examples in **Figure 4.1** and **4.2**. Simulations are used to verify whether Bayesian type I and II error rates, as well as stopping probabilities are the same as desired frequentist values. Let $\widetilde{\alpha_j}$ and $\widetilde{\beta_j}$ denote type I and II error rates at the *j*th analysis. The overall type I and II rates are denoted as $\tilde{\alpha}$ and $\tilde{\beta}$, that $\tilde{\alpha} = \sum_j^3 \widetilde{\alpha_j}$ and $\tilde{\beta} = \sum_j^3 \widetilde{\beta_j}$. Let $\widetilde{SP_{Ej}}$ and $\widetilde{SP_{Fj}}$ denote simulated stopping probabilities under the H_1 for efficacy and futility at the *j*th stage, respectively. Note, the actual Bayesian values, especially the early stage stopping probabilities may slightly deviate from the desired frequentist values due to the uncertainty of simulations and rounding up during stopping boundaries and sample size determinations. Results for simulated type I and II error rates based on single efficacy boundaries and binding double boundaries are summarized in **Table 4.1** and **4.2**; results for stopping probabilities are shown in **Table 4.3** and **4.4**.

Table 4.1 shows simulated Bayesian type I and II error rates for each analysis and overall rates are consistently matched with desired frequentist values in the single efficacy boundaries case. Compared to previous methods proposed by Zhu et al. (Zhu & Yu, 2015) and Shi and Yin(Shi & Yin, 2019), our unified approach is insensitive to the prior choice, therefore strictly control both type I and II error rates for Bayesian methods. For binding doublestopping boundaries in **Table 4.2**, results are generally similar to the single efficacy boundaries. The Bayesian type II error rate generated at each analysis is close to the desired value of beta-spending functions, which means error spending function can be implemented to control type II error rate in Bayesian methods.

As shown in **Table 4.3**, frequentist and Bayesian stopping probabilities for efficacy under the H_1 are very close when non-informative priors are used in the single boundary case. Nevertheless, Bayesian stopping probabilities for efficacy can be different from frequentist probabilities when informative prior distributions are applied. This is because that the distribution

of η under the H_1 and stopping boundary values change with respect to prior distributions. As a result, stopping probabilities can also vary in Bayesian methods given different prior distributions. For the double boundaries case in **Table 4.4**, however, Bayesian and frequentist stopping probabilities are similar for all non-informative and informative priors. This is because both Bayesian type I and II error rates spent at each analysis are set to the same as frequentist values. It suggests that the proposed unified approach can produce the same stopping probabilities under the H_1 between frequentist and Bayesian methods when both efficacy and futility boundaries are present.

Finally, the unified approach is applied to a group sequential clinical trial to demonstrate its application. Suppose there is a clinical trial designed to evaluate the therapeutic benefit of the intensive systolic blood-pressure treatment compared to the standard systolic blood-pressure treatment. The primary outcome is the proportion of patients who do not have severe disability or death at 3 months. Consider that the trial has a 1:1 allocation ratio and a single efficacy boundary. To preserve a nominal significance level at the final analysis that is close to that of a fixed-sample design, O'Brien-Fleming type alpha spending function are used to determine efficacy boundary values. Furthermore, assume there are no covariates or interaction terms to be controlled for evaluating the primary outcome. The overall type I error rate is set to be 0.05 and the desired power is 85% for assumed a 20% increase of primary outcome, from 0.4 to 0.6, in the experimental treatment group. We evaluate the efficacy stopping boundary values (c_E), maximum sample sizes (MSS) and expected sample sizes (ESS) under the alternative hypothesis for two-stages (an interim analysis carried out when 1/2 patients are enrolled), three-stages (interim analyses carried out when 1/3 and 1/2 patients are enrolled) and four-stages (interim analyses carried out when 1/3, 1/2 and 2/3 patients are enrolled). Originally, the trial is designed using frequentist group sequential

methods. Bayesian group sequential methods are derived using the unified approach. Bayesian methods are evaluated under three different prior specifications: a non-informative prior ($\text{Beta}(1,1)$ for both treatment arms), an optimistic prior ($\text{Beta}(30, 10)$ for the experimental arm and ($\text{Beta}(20, 20)$ for the control arm) and a pessimistic prior ($\text{Beta}(16, 24)$ for the experimental arm and ($\text{Beta}(20, 20)$ for the control arm).

Corresponding boundary values, sample size results are summarized in **Table 4.5** and boundaries are also plotted in **Figure 4.3**. For boundary values, we use a standard normal scale for frequentist boundaries and posterior probabilities for Bayesian methods. When non-informative prior distributions are used for Bayesian group sequential methods, the Bayesian maximum sample size and expected sample size are the same as frequentist methods. The optimistic prior reduces the maximum sample sizes needed for Bayesian methods compared to frequentist methods, whereas the pessimistic prior does not increase maximum sample sizes much compared to frequentist methods. For three-stages and four-stages scenarios, optimistic priors reduce the expected sample size in Bayesian methods compared to frequentist methods as well. However, for the two-stage scenario, the Bayesian method with the optimistic prior has a larger expected sample size than the frequentist method. This is due to the fact that Bayesian stopping boundaries are very high and the stopping probability is very small at the interim analysis for the two-stage scenario (See **Figure 4.3**). The pessimistic prior requires slightly more expected sample sizes than frequentist methods. Notably, Bayesian stopping boundary values for pessimistic priors are much smaller than those for non-informative priors and optimistic priors (See **Figure 4.3**). Overall, optimistic priors do not always reduce the expected sample size in the unified approach and pessimistic priors may not increase the maximum sample size and expected sample size much. Clinical trialists should choose from frequentist and Bayesian methods by carefully evaluating the

operating characteristics, e.g. the number of interim analyses and Bayesian priors, when designing a group-sequential trial.

4.4 Conclusion

In this chapter, a novel unified approach for frequentist and Bayesian methods in one-sided two-arm group sequential clinical trials with binary outcomes is developed. The approach unifies frequentist and Bayesian group sequential methods on type I and II error rates. Additionally, the approach unifies stopping probabilities for frequentist and Bayesian group sequential methods with double boundaries. It is assumed that beta conjugate priors are used for treatment arms, and Bayesian decisions depend on posterior probabilities of the difference between the experimental arm and the control arm. The approach utilizes relationships among type I and II error rates, the stopping boundaries, and multivariate distributions of test statistics under the null and alternative hypotheses to achieve unification. Those relationships are well studied in classical group sequential methods(Jennison & Turnbull, 2000). Therefore, the unified approach is intuitive and can be applied in group sequential clinical trials. The approach also helps with the uptake of Bayesian methods into clinical trials as it reduces the computation and generates the same type I and II error rates as frequentist methods do.

4.5 Appendix

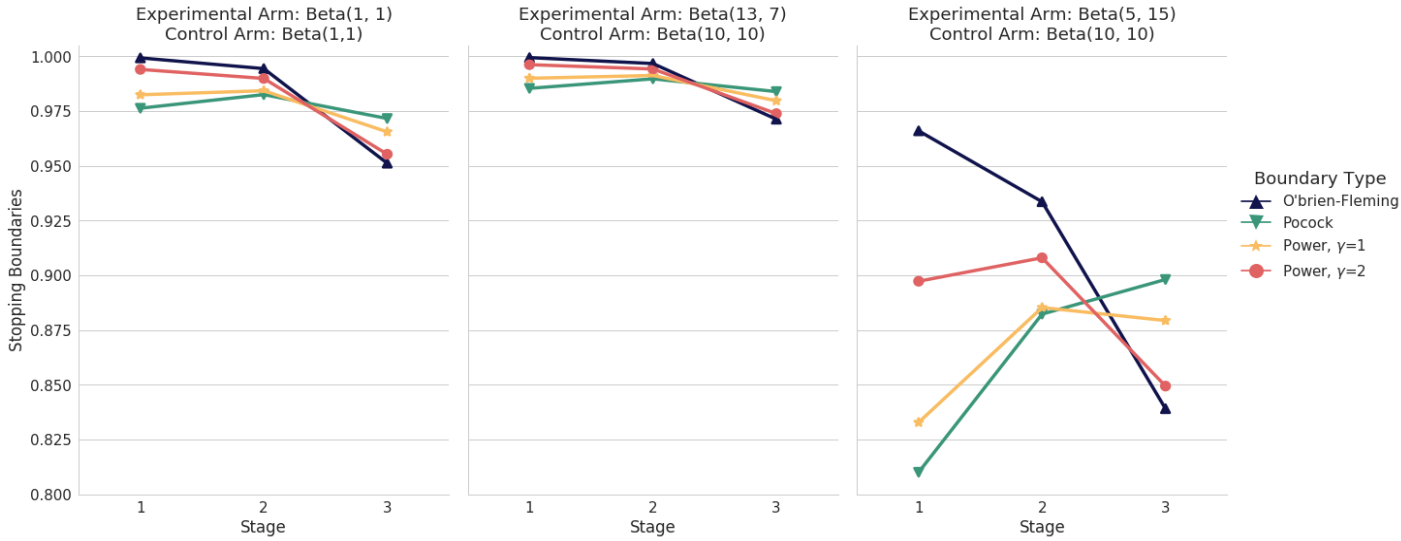


Figure 4.1 Bayesian single efficacy boundary values under three different prior specifications. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$.

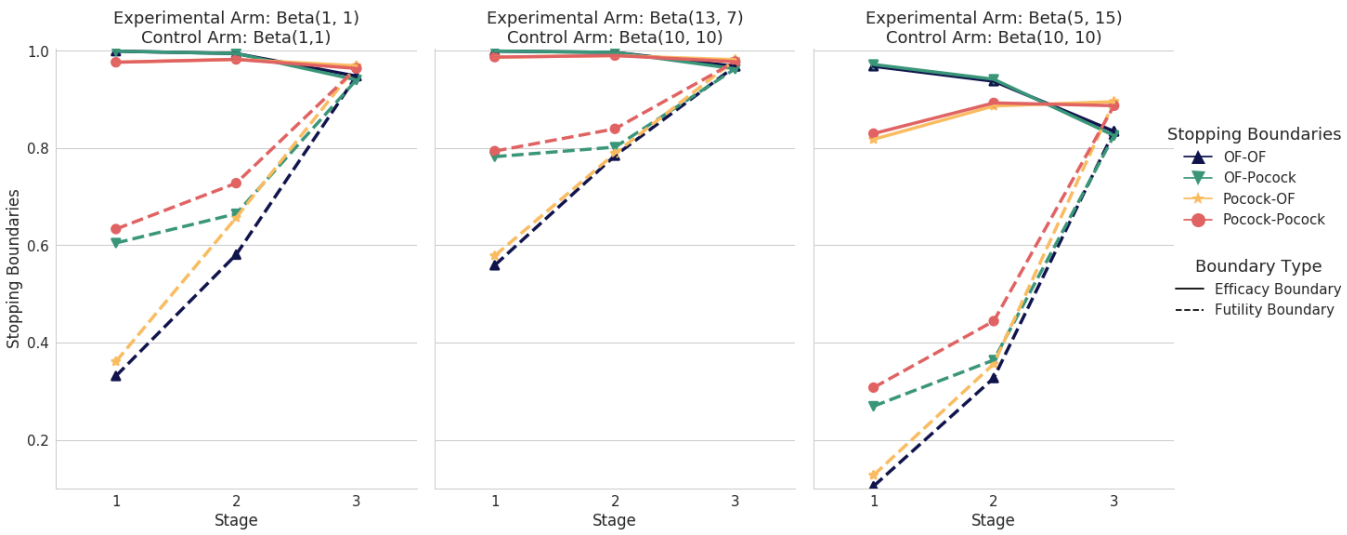


Figure 4.2 Bayesian double boundaries values under three different prior specifications. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$.

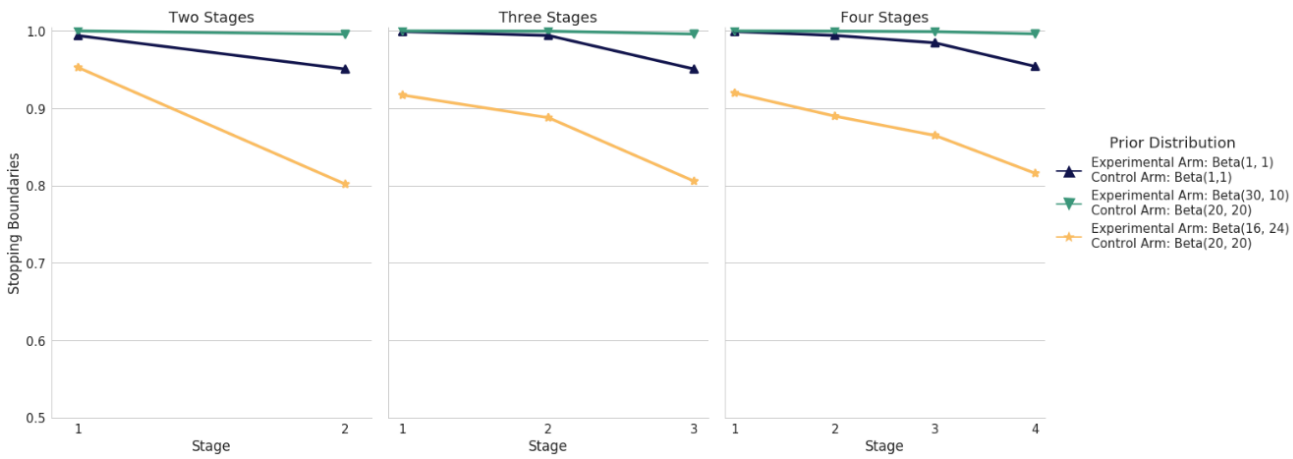


Figure 4.3 Bayesian efficacy stopping boundary values for a specified number of stages and prior distribution in the trial example. It is assumed that $P_1 = 0.6$ and $P_2 = 0.4$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.15$. O'Brien-Fleming type alpha-spending function are used for determining stopping boundaries.

Table 4.1 Type I and II error rates for frequentist and Bayesian group sequential methods with a single efficacy boundary. Bayesian type I and II error rates are simulated under three different prior specifications. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$.

Boundaries	Prior for Treatment	Prior for Control	$\widetilde{\alpha_1}$	$\widetilde{\alpha_2}$	$\widetilde{\alpha_3}$	$\tilde{\alpha}$	$\tilde{\beta}$
O'Brien-Fleming	Desired Frequentist Values		0.000687	0.00489	0.0444	0.05	0.2
	$\text{Beta}(1,1)$	$\text{Beta}(1,1)$	0.00072	0.00434	0.044	0.0491	0.196
	$\text{Beta}(13,7)$	$\text{Beta}(10,10)$	0.000878	0.00498	0.0439	0.0481	0.209
	$\text{Beta}(5,15)$	$\text{Beta}(10,10)$	0.0008	0.0051	0.0460	0.0519	0.205
Pocock	Desired Frequentist Values		0.0226	0.00836	0.019	0.05	0.2
	$\text{Beta}(1,1)$	$\text{Beta}(1,1)$	0.0273	0.0075	0.0184	0.053	0.2
	$\text{Beta}(13,7)$	$\text{Beta}(10,10)$	0.0218	0.0099	0.0165	0.0482	0.205
	$\text{Beta}(5,15)$	$\text{Beta}(10,10)$	0.0207	0.0088	0.0224	0.0519	0.198
Power, $\theta = 1$	Desired Frequentist Values		0.0167	0.00833	0.025	0.05	0.2
	$\text{Beta}(1,1)$	$\text{Beta}(1,1)$	0.0161	0.00911	0.024	0.0492	0.2
	$\text{Beta}(13,7)$	$\text{Beta}(10,10)$	0.0160	0.0103	0.024	0.0503	0.21
	$\text{Beta}(5,15)$	$\text{Beta}(10,10)$	0.0177	0.00662	0.0257	0.050	0.197
Power, $\theta = 2$	Desired Frequentist Values		0.00556	0.00694	0.0375	0.05	0.2
	$\text{Beta}(1,1)$	$\text{Beta}(1,1)$	0.00526	0.00762	0.0344	0.0473	0.196
	$\text{Beta}(13,7)$	$\text{Beta}(10,10)$	0.00644	0.0066	0.0421	0.0551	0.201
	$\text{Beta}(5,15)$	$\text{Beta}(10,10)$	0.00456	0.00928	0.0337	0.0475	0.212

Table 4.2 Type I and II error rates for frequentist and Bayesian group sequential methods with binding double boundaries. Bayesian type I and II error rates are simulated under three different prior specifications. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$.

Boundaries	Prior for Treatment	Prior for Control	$\widetilde{\alpha_1}$	$\widetilde{\alpha_2}$	$\widetilde{\alpha_3}$	$\tilde{\alpha}$	$\widetilde{\beta_1}$	$\widetilde{\beta_2}$	$\widetilde{\beta_3}$	$\tilde{\beta}$
O'Brien-Fleming – O'Brien-Fleming	Desired Frequentist Values		0.000687	0.00489	0.0444	0.05	0.0264	0.0435	0.131	0.2
	$Beta(1,1)$	$Beta(1,1)$	0.000788	0.00522	0.0436	0.0496	0.0264	0.0445	0.136	0.207
	$Beta(13,7)$	$Beta(10,10)$	0.0008	0.00468	0.0434	0.0489	0.0231	0.0563	0.127	0.207
	$Beta(5,15)$	$Beta(10,10)$	0.0008	0.00526	0.0456	0.0517	0.022	0.0443	0.116	0.182
O'Brien-Fleming – Pocock	Desired Frequentist Values		0.000687	0.00489	0.0444	0.05	0.0906	0.0335	0.0759	0.2
	$Beta(1,1)$	$Beta(1,1)$	0.0006	0.00532	0.046	0.0519	0.101	0.0305	0.0732	0.205
	$Beta(13,7)$	$Beta(10,10)$	0.0007	0.00566	0.0431	0.0495	0.101	0.0251	0.0768	0.203
	$Beta(5,15)$	$Beta(10,10)$	0.00058	0.00532	0.046	0.0519	0.101	0.0315	0.0733	0.206
Pocock – O'Brien-Fleming	Desired Frequentist Values		0.0226	0.00836	0.019	0.05	0.0264	0.0435	0.131	0.2
	$Beta(1,1)$	$Beta(1,1)$	0.0213	0.00808	0.0211	0.0505	0.0308	0.0394	0.123	0.194
	$Beta(13,7)$	$Beta(10,10)$	0.0213	0.00808	0.0211	0.0505	0.0308	0.0394	0.123	0.194
	$Beta(5,15)$	$Beta(10,10)$	0.0226	0.00876	0.0198	0.0512	0.029	0.0386	0.118	0.182
Pocock – Pocock	Desired Frequentist Values		0.0226	0.00836	0.019	0.05	0.0906	0.0335	0.0759	0.2
	$Beta(1,1)$	$Beta(1,1)$	0.0279	0.00624	0.0216	0.0558	0.0826	0.0341	0.0635	0.181
	$Beta(13,7)$	$Beta(10,10)$	0.0246	0.0092	0.0181	0.0529	0.107	0.0349	0.0616	0.204
	$Beta(5,15)$	$Beta(10,10)$	0.0276	0.00574	0.0212	0.0545	0.0819	0.0355	0.071	0.188

Table 4.3 Simulated stopping probabilities for frequentist and Bayesian group sequential methods with a single efficacy boundary. Bayesian type I and II error rates are simulated under three different prior specifications. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$.

Boundaries	Prior for Treatment	Prior for Control	$\widetilde{SP_{E1}}$	$\widetilde{SP_{E2}}$	$\widetilde{SP_{E3}}$
O'Brien-Fleming	Frequentist Method		0.0402	0.175	0.582
	Beta(1, 1)	Beta(1,1)	0.040	0.176	0.588
	Beta(13, 7)	Beta(10, 10)	0.0409	0.17	0.580
	Beta(5, 15)	Beta(10, 10)	0.0394	0.168	0.598
Pocock	Frequentist Method		0.343	0.139	0.314
	Beta(1, 1)	Beta(1,1)	0.347	0.138	0.311
	Beta(13, 7)	Beta(10, 10)	0.341	0.15	0.305
	Beta(5, 15)	Beta(10, 10)	0.299	0.142	0.361
Power, $\theta = 1$	Frequentist Method		0.262	0.168	0.369
	Beta(1, 1)	Beta(1,1)	0.253	0.169	0.369
	Beta(13, 7)	Beta(10, 10)	0.253	0.170	0.366
	Beta(5, 15)	Beta(10, 10)	0.269	0.126	0.408
Power, $\theta = 2$	Frequentist Method		0.139	0.166	0.487
	Beta(1, 1)	Beta(1,1)	0.137	0.168	0.5
	Beta(13, 7)	Beta(10, 10)	0.149	0.158	0.492
	Beta(5, 15)	Beta(10, 10)	0.115	0.201	0.473

Table 4.4 Simulated stopping probabilities for frequentist and Bayesian group sequential methods with binding double boundaries. Bayesian type I and II error rates are simulated under three different prior specifications. It is assumed that $P_1 = 0.65$ and $P_2 = 0.5$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.2$.

Boundaries	Prior for Treatment	Prior for Control	$\widetilde{SP_{E1}}$	$\widetilde{SP_{E2}}$	$\widetilde{SP_{E3}}$	$\widetilde{SP_{F1}}$	$\widetilde{SP_{F2}}$	$\widetilde{SP_{F3}}$
O'Brien-Fleming – O'Brien-Fleming	Frequentist Method		0.304	0.316	0.332	0.0258	0.0435	0.135
	$Beta(1,1)$	$Beta(1,1)$	0.302	0.313	0.335	0.0264	0.0445	0.136
	$Beta(13,7)$	$Beta(10,10)$	0.306	0.345	0.3	0.0231	0.0563	0.127
	$Beta(5,15)$	$Beta(10,10)$	0.303	0.312	0.33	0.022	0.0443	0.116
O'Brien-Fleming – Pocock	Frequentist Method		0.618	0.111	0.221	0.1	0.0294	0.0744
	$Beta(1,1)$	$Beta(1,1)$	0.614	0.111	0.222	0.101	0.0305	0.0732
	$Beta(13,7)$	$Beta(10,10)$	0.619	0.111	0.221	0.101	0.0251	0.0768
	$Beta(5,15)$	$Beta(10,10)$	0.614	0.111	0.222	0.101	0.0315	0.0733
Pocock – O'Brien-Fleming	Frequentist Method		0.38	0.283	0.282	0.0264	0.0403	0.128
	$Beta(1,1)$	$Beta(1,1)$	0.386	0.282	0.282	0.0308	0.0394	0.123
	$Beta(13,7)$	$Beta(10,10)$	0.386	0.282	0.282	0.0308	0.0394	0.123
	$Beta(5,15)$	$Beta(10,10)$	0.386	0.28	0.283	0.029	0.0386	0.118
Pocock – Pocock	Frequentist Method		0.613	0.175	0.163	0.0914	0.0336	0.0649
	$Beta(1,1)$	$Beta(1,1)$	0.611	0.170	0.162	0.0826	0.0341	0.0635
	$Beta(13,7)$	$Beta(10,10)$	0.617	0.173	0.157	0.107	0.0349	0.0616
	$Beta(5,15)$	$Beta(10,10)$	0.610	0.162	0.173	0.0819	0.0355	0.071

Table 4.5 Efficacy stopping boundary values (c_E), Maximum sample sizes (MSS) and expected sample sizes (ESS) under the alternative hypothesis for example trial designed with frequentist and Bayesian group sequential methods with a different number of stages. It is assumed that $P_1 = 0.6$ and $P_2 = 0.4$, $r = 1$, $\Delta = 0$, $\alpha = 0.05$ and $\beta = 0.15$. O'Brien-Fleming type alpha-spending function are used for determining stopping boundaries.

Design			Frequentist Methods	Bayesian Methods		
				Experimental: Beta(1,1) Control: Beta(1,1)	Experimental: Beta(30,10) Control: Beta(20,20)	Experimental: Beta(16,24) Control: Beta(20,20)
Two-Stages	c_E	Stage 1	2.538	0.994	0.999	0.953
		Stage 2	1.662	0.951	0.996	0.802
	MSS		174	174	164	176
	ESS		152	152	155	168
Three-Stages	c_E	Stage 1	3.202	0.999	0.999	0.917
		Stage 2	2.552	0.994	0.999	0.888
		Stage 3	1.662	0.951	0.996	0.806
	MSS		174	174	164	176
	ESS		150	150	139	152
Four-Stages	c_E	Stage 1	3.202	0.999	0.999	0.920
		Stage 2	2.552	0.994	0.999	0.890
		Stage 3	2.179	0.985	0.999	0.865
		Stage 4	1.698	0.954	0.996	0.816
	MSS		178	178	166	178
	ESS		136	136	127	138

CHAPTER FIVE: SPECIFIC AIM 3

5.1 Introduction

As mentioned in **Chapter 1**, there is a long-time debate between frequentist and Bayesian clinical trialist on which approach is better. Rather than declaring one to be better than the other, Bayarri and Berger(Bayarri & Berger, 2004) had acknowledged that frequentist and Bayesian approaches can both play important roles in clinical trials. In fact, only a few parallel comparisons are made between frequentist and Bayesian approaches because of a lack of established correspondence between two paradigms (Inoue et al., 2005; Lewis et al., 2007). To be more specific, frequentist and Bayesian clinical trialists may claim one paradigm is superior to the other, however, the targeted operating characteristics of two paradigms are quite different. Additionally, it should be noticed that many of the 'frequentist vs Bayesian comparisons were carried out from the Bayesian clinical trialists' side while few came from frequentist perspectives. Overall, a fair comparison should be made between frequentist and Bayesian approaches, while holding the operating characteristics such as type I and II error rates as the same.

Considering the above issues, two novel unified approaches for frequentist and Bayesian methods in fixed-sample and group-sequential clinical trials with binary endpoints are proposed in **Chapter 3** and **Chapter 4**. The unified approaches allowed Bayesian and frequentist approaches to generate the same type I and II error rates. For Bayesian approaches, there is no need to adjust the prior distributions, but Bayesian sample sizes and decision thresholds are calibrated to satisfy type I and II error rates. Hence, a detailed numerical investigation can be nicely performed to compare Bayesian and frequentist approaches in the unified framework for different design parameters, such as Bayesian prior specifications, numbers of analyses, allocation ratios and stopping boundaries.

In this chapter, a deep investigation, motivated by a phase III trial for hemorrhagic stroke is conducted. First, the influences of different prior specifications and design parameters on the Bayesian maximum and expected sample sizes and decision thresholds are evaluated. Second, frequentist and Bayesian approaches are studied in a comparable manner. Finally, suggestions for making a selection between Bayesian and frequentist approaches for specific clinical trial designs are given. **Section 5.2** briefly introduced the methods and **Section 5.3** summarizes the results.

5.2 Methods

Consider the same two-arm phase III randomized stroke trial as the one used in a trial application in **Chapter 4**. The trial evaluates the superiority of the intensive systolic blood-pressure reduction treatment compared to the standard systolic blood-pressure reduction treatment after the onset of the hemorrhagic stroke. The primary endpoint is the proportion who do not have severe disability or death at 3 months and is binomially distributed. Patients are randomized into two treatment arms. The type I error rate and power are specified to be 0.05 and 0.85. It is assumed that the intensive systolic blood-pressure would increase the proportion of patients who satisfy the primary endpoint by 20%, from 0.4 to 0.6. In addition, it is supposed that there are no covariates or interaction terms when analyzing the primary endpoint. Originally, the trial is designed using frequentist approaches. However, it is re-designed and evaluated using Bayesian approaches, in order to compare Bayesian and frequentist sample sizes and decision thresholds under various conditions. It is of particular interest to compare Bayesian and frequentist properties with different prior specifications, allocation ratios, and the number of analyses.

The frequentist methods and Bayesian methods are derived for specific trial parameters using the unified approaches proposed in **Chapter 3** and **4**. Trial parameters include the Bayesian priors, the allocation ratio and the number of analyses. There are many prior elicitation methods

available, e.g.(Fox, 1966; Gross, 1971; Ibrahim & Chen, 2000; Wu, Shih, & Moore, 2008). An intuitive approach is used here to specify the prior distribution(Gross, 1971), in which prior mean centers at the prior belief about the treatment effect. Prior sample size proposed by Morita et al. (Morita et al., 2012) is used to denote the sum of prior parameters, that is, $a + b$. The larger the prior sample size is, the greater effect the prior distribution can exert on the posterior distribution. To evaluate the effect of prior sample size on Bayesian sample sizes, a range of $a + b$ values are considered, while holding $a/(a + b)$ to be fixed.

Notably, prior distributions are categorized into three classes, the non-informative priors, the optimistic priors (or the enthusiastic priors) and the pessimistic priors (or the skeptical priors), based on the prior mean (Moatti et al., 2013). The non-informative prior minimizes the influence of prior on the posterior distribution and let data contributes most to the posterior distribution. The optimistic prior considers a beneficial treatment effect and the prior mean is greater than or equal the assumed treatment effect of the intensive systolic blood pressure reduction arm. The pessimistic prior considers a treatment effect is unlikely to be observed, and the prior mean is no greater than the assumed treatment effect of the standard systolic blood pressure reduction arm.

In this chapter, influences of Bayesian priors are first evaluated using a $Beta(1,1)$ non-informative prior distribution. Then both optimistic and pessimistic priors are applied to the intensive systolic blood pressure reduction treatment group. Optimistic priors with mean values of 0.8 and 0.6 are considered in the evaluation, while for pessimistic priors, mean values of 0.4 and 0.2 are considered. For the standard treatment group, only pessimistic priors with a prior mean centered on the null hypothesis, i.e. $a/(a + b) = 0.4$ are considered.

All evaluations are conducted using a user-friendly software application. The application has been developed and made accessible online to allow other clinical trialists to use the unified

framework when designing their clinical trials. PyQt framework is used to develop the frontend of the application and all algorithms are implemented using Cython. The interface of the application is presented in **Figure 5.1**. Users can simply enter the parameters in the input box, click and run to get the results. Note, the application also supports simulating operating characteristics for Bayesian multi-arm multi-stage design (Yu et al., 2019) using an optimized algorithm proposed earlier. Full documentation of the software application can be found on the website http://usebats.org/bats.

5.3 Results

Bayesian sample sizes and decision thresholds are evaluated for five prior specifications mentioned above (prior mean of 0.8, 0.6, 0.4, 0.2 and the non-informative prior), with informative prior sample sizes ranging from 10 to 60 by 10. Here, prior sample sizes for the experimental arm and the control arm are held the same. For simplicity, it is assumed that patients are equally randomized to the experimental arm and the control arm. Further, assume there are two interim analyses planned when 1/3 and 1/2 patients are enrolled and only treatment efficacy is assessed at interim analyses. An O'Brien-Fleming type alpha spending function has been used to derive the efficacy boundaries. The type I error rate and power are specified to be 0.05 and 0.85, as stated above. Bayesian maximum sample sizes and expected sample sizes are plotted against the prior sample sizes in **Figure 5.2**. Bayesian stopping boundary values are shown in **Figure 5.3**. Corresponding frequentist maximum sample sizes, expected sample sizes and stopping boundaries are also calculated. Complete results are summarized in tables in **Table 5.1**.

Figure 5.2 displays the trends of Bayesian maximum and expected sample sizes for corresponding prior sample sizes. The dashed line represents the maximum sample size and the expected sample size for the $Beta(1,1)$ non-informative prior, which coincide with the frequentist

values. The maximum sample size reduces significantly for the optimistic prior with a mean of 0.8. As the prior sample size $a + b$ increases from 10 to 60, the maximum sample size decreases from 170 to 150. Nevertheless, the maximum sample size does not change much for the other optimistic prior with a mean of 0.6. For two pessimistic priors, the maximum sample sizes are nearly the same as the maximum sample size for the non-informative prior. Similar patterns are observed for the expected sample size. The optimistic prior with a mean of 0.8 has the largest expected sample size reduction, while the other optimistic prior results in a negligible expected sample size reduction. Expected sample sizes for two pessimistic priors are slightly greater than that for the non-informative prior.

Notably, stopping boundaries are also adjusted (see **Figure 5.3**). Stopping boundary values for optimistic priors are more rigorous than those for the non-informative prior. An optimistic prior with large prior sample size can produce rigid stopping rules across all stages. Conversely, stopping boundary values for pessimistic priors decrease as the prior sample size increases, and are smaller than stopping boundaries for the non-informative prior. Maximum sample size

Next, allocation ratios $r = 1,2,3$ are examined for the five prior specifications, assuming the same three-stage design with an O'Brien-Fleming efficacy boundary as above. In addition, consider prior sample sizes of 5, 7.5, 15, 30 and 45 for the intensive systolic blood pressure reduction group. A prior sample size of 15 is fixed for the standard treatment group. **Figure 5.4** and 5.**5** present heatmaps of the maximum and expected sample sizes for each prior specification, the x axis represents the allocation ratios and the y axis represents the prior sample sizes for the experimental arm. The lighter the color of the cell in the heatmap, the larger the sample size is. The maximum sample size, expected sample size and stopping boundaries for the non-informative

prior case and the frequentist approach are summarized in **Table 5.2**. Complete results including stopping boundary values are included in **Table 5.3**.

When non-informative priors are applied to treatment arms, the Bayesian maximum and expected sample sizes are similar to frequentist values if patients are equally allocated to two treatment arms (see **Table 5.2**). However, for unequal allocation ratios, Bayesian maximum and expected sample sizes are smaller than frequentist values. Stopping boundary values do not change much with respect to allocation ratio changes. The maximum sample size increases as the allocation ratio become large. In contrast to the non-informative priors, the maximum sample size is constantly the smallest for all informative priors when allocation ratio $r = 2$ (see **Figure 5.4**). While the maximum sample sizes are roughly the same between $r = 1$ and $r = 3$ the prior sample size of the experimental arm is smaller than that of the control arm. Notably, the ratio of prior sample sizes can affect the maximum sample size value. The maximum sample size increases as the ratio of the prior sample sizes of the experimental arm to the control arm increases.

As such, the expected sample sizes are the smallest for allocation ratio $r = 2$ for all informative prior. The expected sample sizes for $r = 3$ is slightly smaller than that for $r = 1$. Compared to frequentist approaches, Bayesian expected sample sizes can be less, equal or more, depending on the allocation ratio and prior sample size ratio. For equal allocation, Bayesian expected sample sizes are smaller than frequentist expected sample sizes, when optimistic priors are used and prior sample sizes of the experimental arm are smaller than those of the control arm. For $r = 2$, Bayesian expected sample sizes are smaller than frequentist expected sample sizes, as long as prior sample sizes of the experimental arm are smaller than those of the control arm. While for $r = 3$, the Bayesian expected sample sizes are consistently smaller than the frequentist expected

sample sizes. Results indicates the allocation ratio, as well as the ratio of prior sample sizes, can influence the Bayesian sample sizes.

Further, the influences of the number of analyses on Bayesian maximum and expected sample sizes are assessed. Assume that patients are equally randomized into two treatment arms and the prior sample sizes are set to 30 for both treatment arms. The same type I and II error rates as above are used. Multiple numbers of analyses ranging from 1 to 6 are evaluated. Further, it is assumed that there is equal spacing between two scheduled analyses and O'Brien-Fleming efficacy boundary is used. **Figure 5.6** shows the Bayesian maximum and expected sample sizes plotted against a various number of analyses. Note, planning only one analysis leads to a fixed-sample one-stage trial. Hence, the expected sample size is reported as the maximum sample size. Corresponding frequentist maximum sample sizes, expected sample sizes and stopping boundaries are also derived. Complete results are summarized in tables in **Table 5.4**.

From the results, it can be found that the Bayesian maximum sample size increases as the number of analyses increases, while the expected sample size becomes smaller when there are more interim analyses. Bayesian maximum and expected sample sizes are close to frequentist values when non-informative prior is used. Similar to results of evaluations for prior sample sizes, applying optimistic priors lead to smaller maximum and expected sample sizes than corresponding frequentist sample sizes, whereas pessimistic prior lead to larger sample size values.

5.4. Conclusion

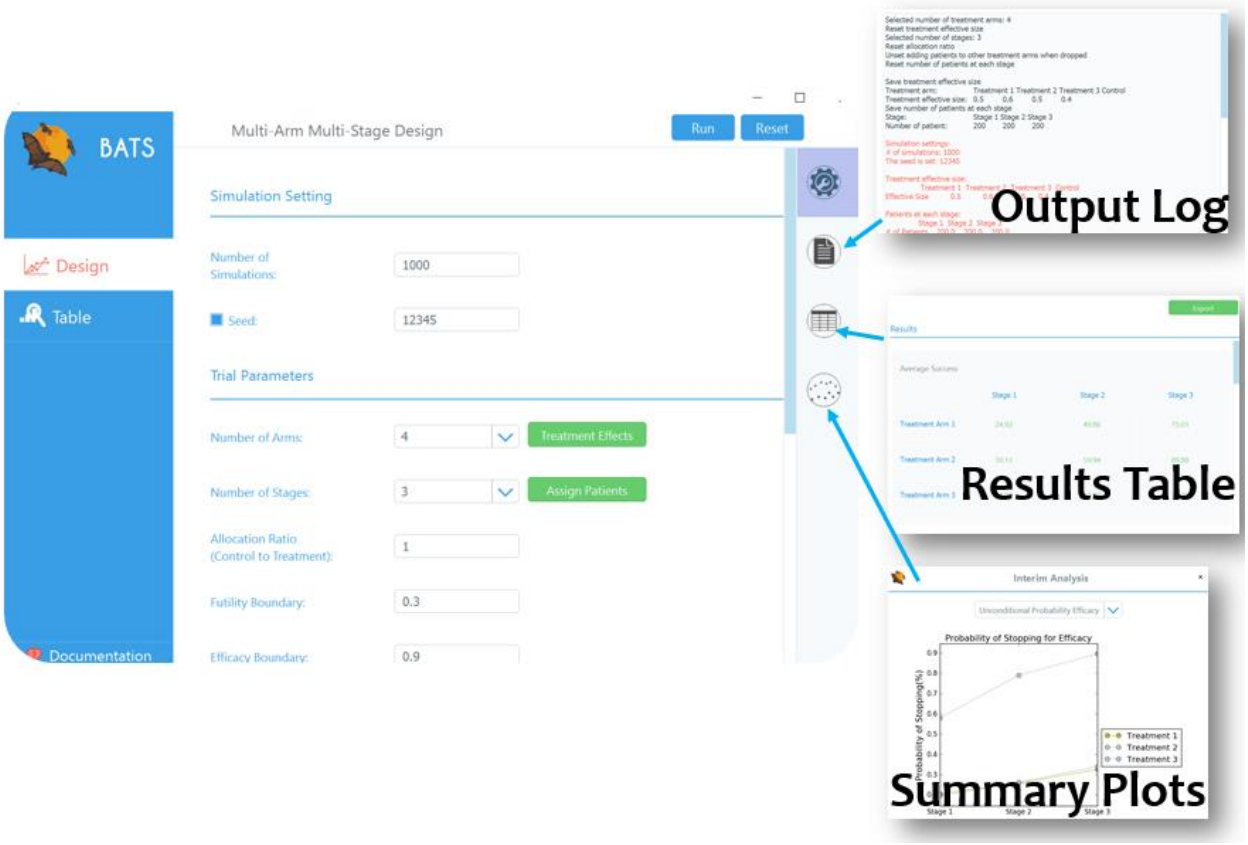
The Bayesian-frequentist debate in clinical trials has been there for a long time (Efron, 2005; David J Spiegelhalter, Freedman, & Parmar, 1994b). For Bayesian methods, the use of subjective Bayesian priors is also of concern to some clinical trialists. Rather than making assertive statements on problems, like which statistical paradigm is better, or whether Bayesian prior

information is good or not, this chapter focuses on evaluating these problems under different circumstances. An investigation is present to assess Bayesian prior influences on the maximum and expected sample sizes, decision thresholds, with varying prior sample sizes, allocation ratios and number of analyses. Meanwhile, comparisons between sample sizes of Bayesian approaches and sample sizes of corresponding frequentist approaches are made. The comparisons are objective because compared Bayesian and frequentist methods lead to the same conclusions.

Results show that selection of different Bayesian priors, prior sample sizes, allocation ratios and a number of analyses can change Bayesian maximum and expected sample sizes. Comparisons between Bayesian and frequentist results show that when non-informative priors are applied to treatment arms in Bayesian approaches, resulting in maximum and expected sample sizes are the same as frequentist values for most cases. Lewis et al. (Lewis et al., 2007) made a similar conclusion in a previous paper, but the authors compared a Bayesian decision-theoretical approach with the classical frequentist approach. Only exceptions are when unequal allocation ratios are used, indicating that Bayesian methods may be beneficial for sample size reduction in clinical trials with unequal allocations.

5.5 Appendix

Figure 5.1 The interface of the software application.



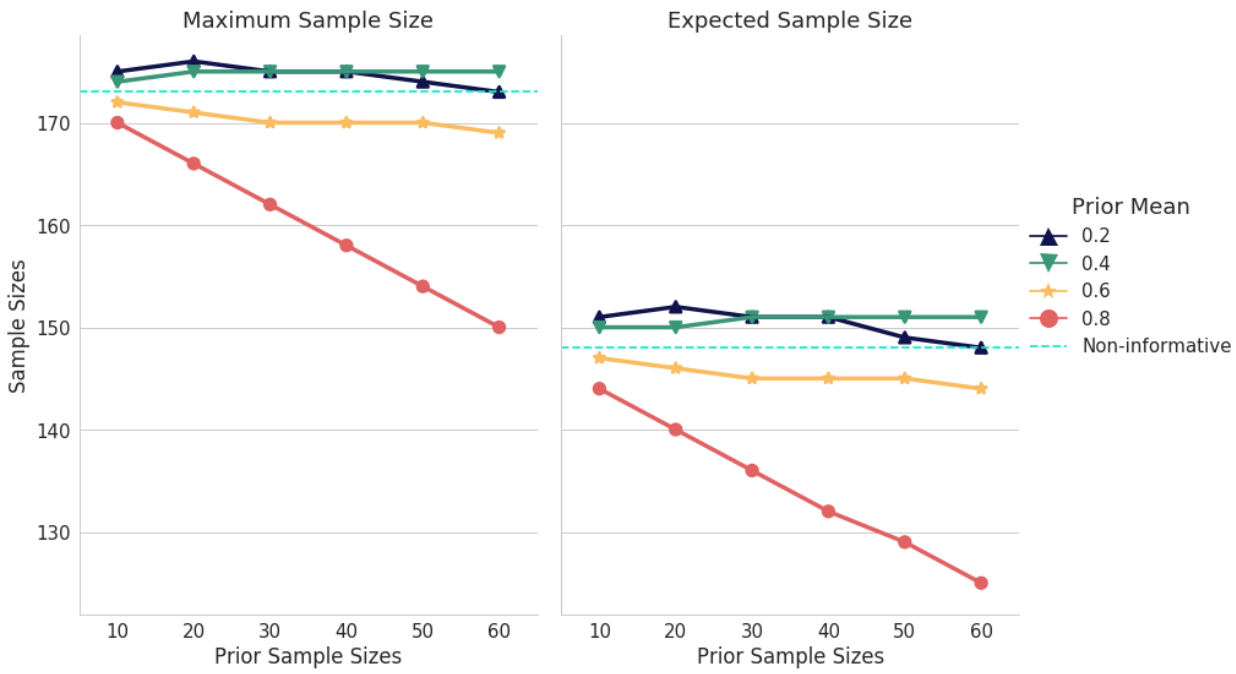


Figure 5.2 Bayesian maximum and expected sample sizes with respect to prior sample sizes. Light blue dashed lines represent the maximum and expected sample sizes when non-informative prior is used.

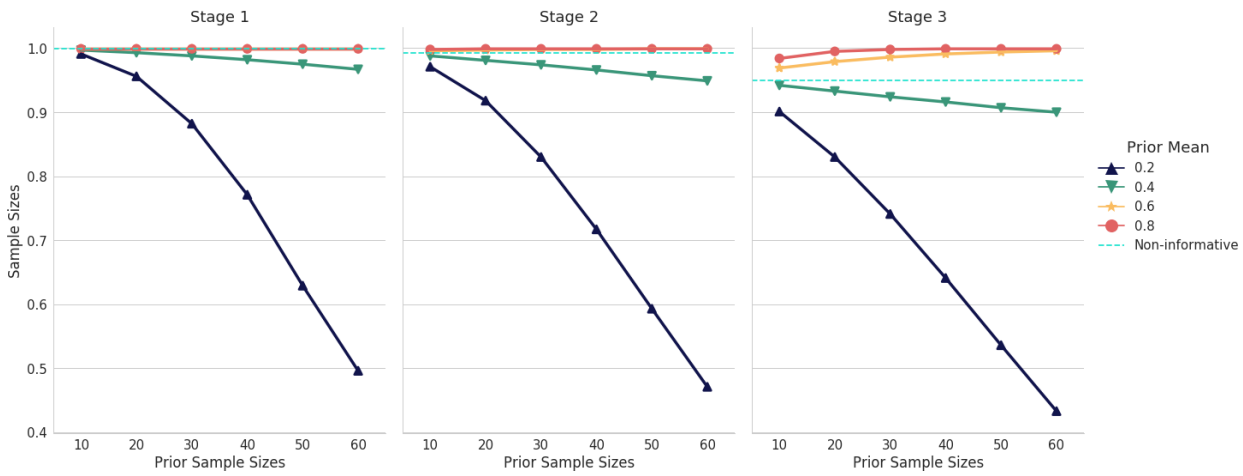


Figure 5.3 Bayesian stopping boundary (decision threshold) value at each stage for different prior sample sizes. Light blue dashed lines represent stopping boundary values when non-informative prior is used.

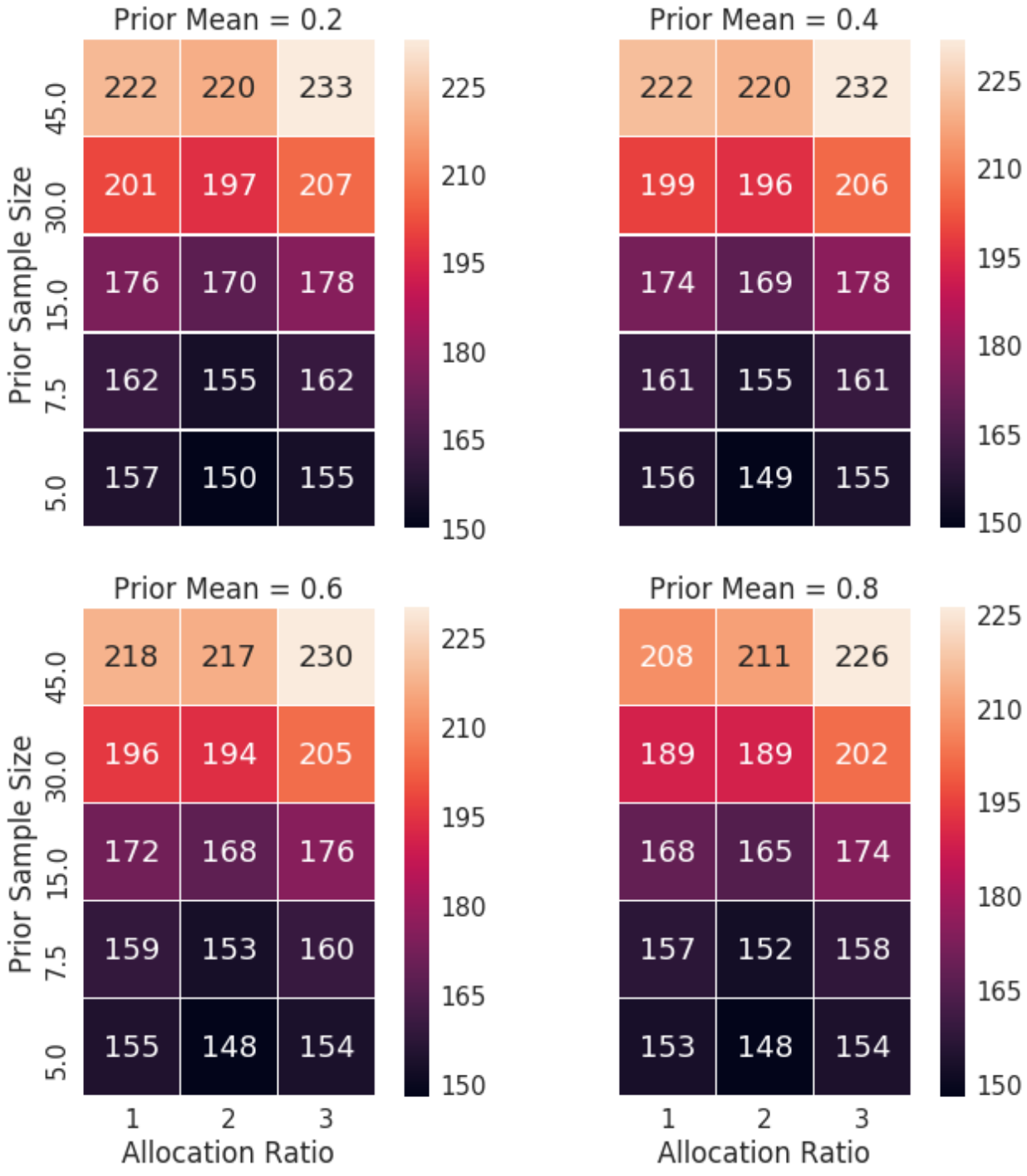


Figure 5.4. Bayesian maximum sample sizes with different allocation ratios and prior sample sizes for the intensive systolic blood pressure reduction treatment.

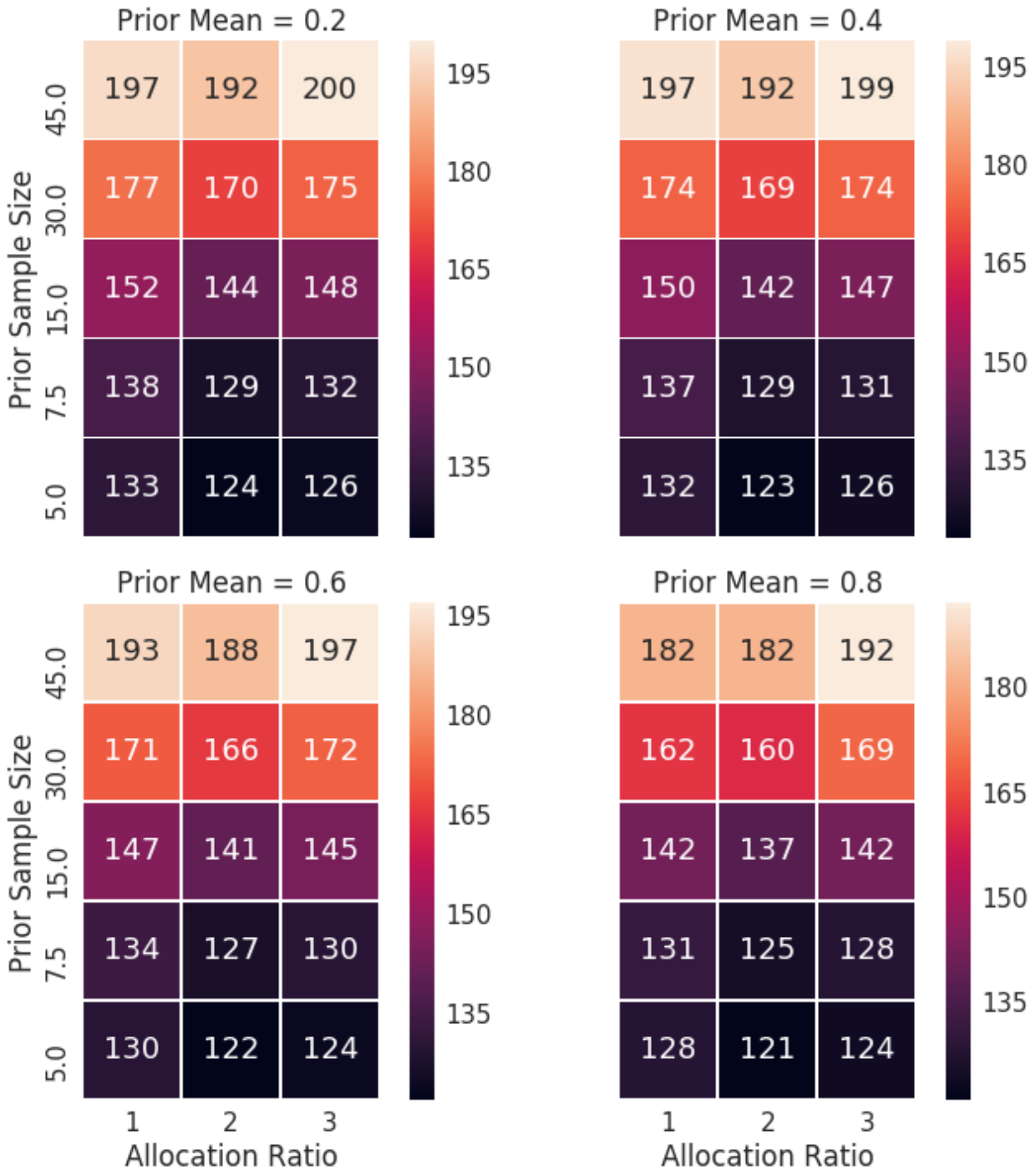


Figure 5.5. Bayesian expected sample sizes with different allocation ratios and prior sample sizes for the intensive systolic blood pressure reduction treatment.

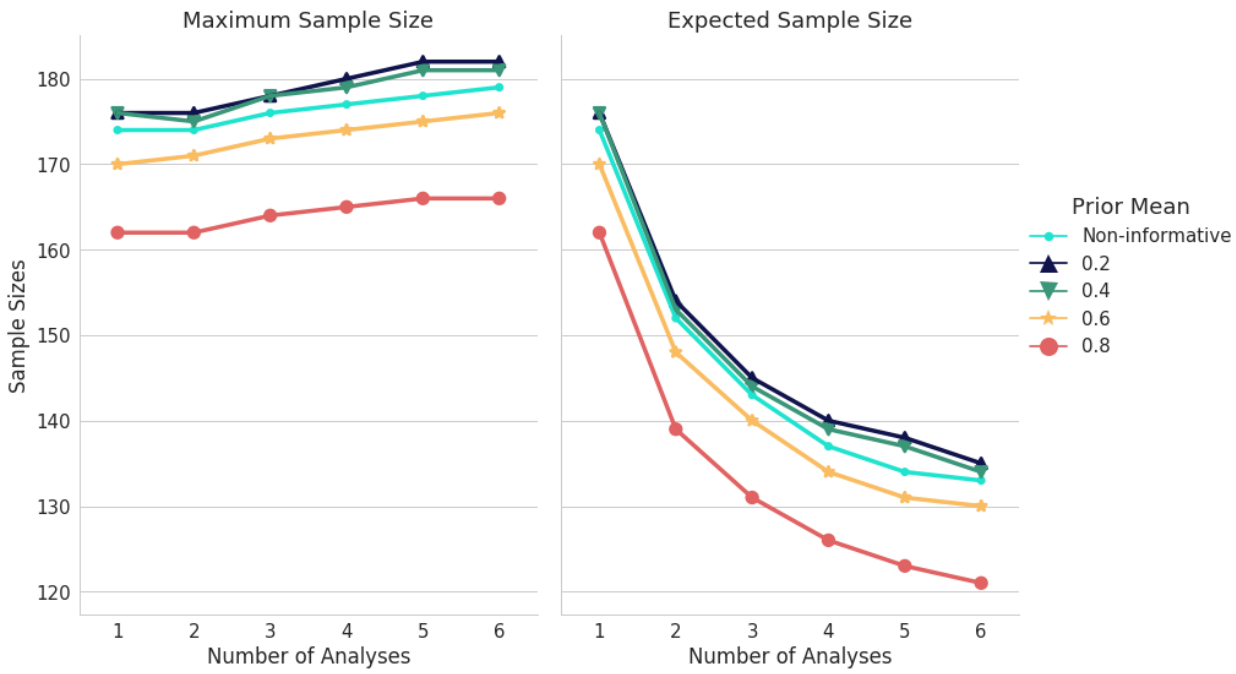


Figure 5.6. Bayesian maximum and expected sample sizes with respect to the number of analyses. Light blue dashed lines represent the maximum and expected sample sizes when non-informative prior is used.

Table 5.1 Complete results of maximum sample size (MSS), expected sample size (ESS) and decision thresholds for Bayesian methods when changing Bayesian prior sample sizes. Values for frequentist approaches are also reported.

Prior Mean	Prior Sample Size	MSS	ESS	Stopping Boundaries		
				Stage 1	Stage 2	Stage 3
0.8	10	170	144	0.999	0.998	0.984
	20	166	140	0.999	0.999	0.995
	30	162	136	0.999	0.999	0.998
	40	158	132	0.999	0.999	0.999
	50	154	129	0.999	0.999	0.999
	60	150	125	0.999	0.999	0.999
0.6	10	172	147	0.999	0.996	0.969
	20	171	146	0.999	0.997	0.979
	30	170	145	0.999	0.998	0.986
	40	170	145	0.999	0.998	0.991
	50	170	145	0.999	0.999	0.994
	60	169	144	0.999	0.999	0.996
0.4	10	174	150	0.997	0.988	0.942
	20	175	150	0.993	0.981	0.933
	30	175	151	0.988	0.974	0.924
	40	175	151	0.982	0.966	0.916
	50	175	151	0.975	0.957	0.907
	60	175	151	0.967	0.949	0.9
0.2	10	175	151	0.991	0.971	0.901
	20	176	152	0.956	0.918	0.83
	30	175	151	0.882	0.83	0.741
	40	175	151	0.771	0.717	0.641
	50	174	149	0.629	0.593	0.536
	60	173	148	0.496	0.471	0.433
Non-informative Prior		173	148	0.999	0.993	0.95
Frequentist		173	148	3.202	2.552	1.662

Table 5.2 The maximum and expected sample sizes and decision thresholds for the Bayesian approach with non-informative priors and the corresponding frequentist approach, evaluating with different allocation ratios.

Parameter		Allocation Ratio					
		$r = 1$		$r = 2$		$r = 3$	
		Bayesian	Frequentist	Bayesian	Frequentist	Bayesian	Frequentist
c_E	Stage 1	0.999	3.202	0.999	3.202	0.999	3.202
	Stage 2	0.993	2.552	0.993	2.552	0.992	2.552
	Stage 3	0.950	1.662	0.948	1.662	0.947	1.662
MSS		174	174	192	197	222	232
ESS		148	150	163	168	189	200

Table 5.3 Complete results of maximum and expected sample sizes and decision thresholds for Bayesian methods when changing allocation ratios and prior sample sizes for the experimental arm. Values for frequentist approaches are also reported.

Prior Sample Size for the Experimental Arm	Prior Mean	Allocation Ratio (The experimental arm to the control)														
		1:1					2:1					3:1				
		MSS	ESS	Stopping Boundaries			MSS	ESS	Stopping Boundaries			MSS	ESS	Stopping Boundaries		
				Stage 1	Stage 2	Stage 3			Stage 1	Stage 2	Stage 3			Stage 1	Stage 2	Stage 3
5	0.8	153	128	0.999	0.996	0.971	148	121	0.999	0.994	0.961	154	124	0.998	0.991	0.954
	0.6	155	130	0.998	0.993	0.958	148	122	0.998	0.99	0.95	154	124	0.997	0.987	0.943
	0.4	156	132	0.997	0.988	0.942	149	123	0.996	0.985	0.936	155	126	0.994	0.982	0.931
	0.2	157	133	0.994	0.98	0.921	150	124	0.993	0.978	0.92	155	126	0.991	0.975	0.917
7.5	0.8	157	131	0.999	0.997	0.978	152	125	0.999	0.996	0.969	158	128	0.998	0.994	0.962
	0.6	159	134	0.999	0.994	0.963	153	127	0.998	0.992	0.955	160	130	0.997	0.989	0.948
	0.4	161	137	0.997	0.987	0.941	155	129	0.995	0.984	0.936	161	131	0.994	0.981	0.931
	0.2	162	138	0.991	0.973	0.908	155	129	0.99	0.972	0.911	162	132	0.988	0.969	0.909
15	0.8	168	142	0.999	0.999	0.991	165	137	0.999	0.998	0.984	174	142	0.999	0.997	0.978
	0.6	172	147	0.999	0.996	0.975	168	141	0.998	0.994	0.966	176	145	0.998	0.992	0.96
	0.4	174	150	0.995	0.985	0.938	169	142	0.994	0.983	0.934	178	147	0.993	0.98	0.931
	0.2	176	152	0.978	0.949	0.868	170	144	0.979	0.954	0.883	178	148	0.978	0.954	0.888
30	0.8	189	162	0.999	0.999	0.998	189	160	0.999	0.999	0.995	202	169	0.999	0.999	0.991
	0.6	196	171	0.999	0.998	0.987	194	166	0.999	0.997	0.98	205	172	0.999	0.996	0.974
	0.4	199	174	0.994	0.983	0.934	196	169	0.993	0.981	0.933	206	174	0.992	0.98	0.931
	0.2	201	177	0.942	0.888	0.781	197	170	0.954	0.917	0.828	207	175	0.957	0.925	0.846
45	0.8	208	182	0.999	0.999	0.999	211	182	0.999	0.999	0.998	226	192	0.999	0.999	0.996
	0.6	218	193	0.999	0.999	0.993	217	188	0.999	0.998	0.987	230	197	0.999	0.997	0.982
	0.4	222	197	0.994	0.982	0.932	220	192	0.993	0.981	0.933	232	199	0.992	0.98	0.931
	0.2	222	197	0.9	0.827	0.696	220	192	0.929	0.877	0.775	233	200	0.938	0.897	0.807
Non-informative		174	148	0.999	0.993	0.95	192	163	0.999	0.993	0.948	222	189	0.999	0.992	0.947
Frequentist		174	150	3.202	2.552	1.662	197	168	3.202	2.552	1.662	232	200	3.202	2.552	1.662

Table 5.4 Complete results of maximum and expected sample sizes and decision thresholds for Bayesian methods when changing number of analyses. Values for frequentist approaches are also reported.

Number of Analyses	Prior Mean	MSS	ESS	Stopping Boundaries					
				Stage 1	Stage 2	Stage 3	Stage 4	Stage 5	Stage 6
1	0.8	162	162	0.999	-	-	-	-	-
	0.6	170	170	0.986	-	-	-	-	-
	0.4	176	176	0.923	-	-	-	-	-
	0.2	176	176	0.738	-	-	-	-	-
	Non-informative Prior	174	174	0.949	-	-	-	-	-
	Frequentist	174	174	1.64	-	-	-	-	-
2	0.8	162	139	0.999	0.998	-	-	-	-
	0.6	171	148	0.998	0.986	-	-	-	-
	0.4	175	153	0.975	0.925	-	-	-	-
	0.2	176	154	0.836	0.741	-	-	-	-
	Non-informative Prior	174	152	0.993	0.95	-	-	-	-
	Frequentist	174	151	2.538	1.662	-	-	-	-
3	0.8	164	131	0.999	0.999	0.998	-	-	-
	0.6	173	140	0.999	0.995	0.987	-	-	-
	0.4	178	144	0.988	0.96	0.929	-	-	-
	0.2	178	145	0.883	0.803	0.752	-	-	-
	Non-informative Prior	176	143	0.999	0.982	0.953	-	-	-
	Frequentist	176	152	3.202	2.142	1.694	-	-	-
4	0.8	165	126	0.999	0.999	0.999	0.999	-	-
	0.6	174	134	0.999	0.998	0.994	0.988	-	-
	0.4	179	139	0.993	0.976	0.954	0.932	-	-
	0.2	180	140	0.905	0.84	0.794	0.761	-	-
	Non-informative Prior	177	137	0.999	0.993	0.977	0.956	-	-
	Frequentist	178	137	3.75	2.54	2.016	1.72	-	-
5	0.8	166	123	0.999	0.999	0.999	0.999	0.999	-
	0.6	175	131	0.999	0.999	0.997	0.993	0.988	-
	0.4	181	137	0.995	0.984	0.968	0.951	0.934	-
	0.2	182	138	0.923	0.866	0.824	0.792	0.768	-
	Non-informative Prior	178	134	0.999	0.997	0.988	0.974	0.957	-
	Frequentist	179	134	4.229	2.889	2.298	1.962	1.74	-
6	0.8	166	121	0.999	0.999	0.999	0.999	0.999	0.999
	0.6	176	130	0.999	0.999	0.998	0.996	0.993	0.989
	0.4	181	134	0.997	0.989	0.976	0.963	0.949	0.936
	0.2	182	135	0.937	0.884	0.846	0.815	0.791	0.772
	Non-informative Prior	179	133	0.999	0.999	0.994	0.984	0.972	0.959
	Frequentist	180	132	4.655	3.202	2.552	2.179	1.934	1.755

CHAPTER SIX: CONCLUSIONS

6.1 Conclusion

In this dissertation, we develop two unified approaches for frequentist and Bayesian hypothesis tests in two-arm fixed-sample and group-sequential superiority trials with binary endpoints. We assume the Bayesian prior follows a conjugated beta distribution and the Bayesian test is based on the posterior probability of rate difference. The idea of the proposed approaches for unifying frequentist and Bayesian methods is new and much improved compared to other proposed methodologies (Shi & Yin, 2019; Zhu & Yu, 2015). First, proposed unified approaches took Bayesian prior distributions into sample size and decision threshold determinations. As different priors can influence stopping boundaries and sample size and resulting type I and II error rates, taking the Bayesian prior into account controls type I and II error rates at the desired level. Second, a theoretical approach to determine Bayesian sample sizes and decision thresholds are provided, which does not require any simulation is very quick to compute. Last, varying binding and non-binding futility boundary were considered in the unified approach in **Chapter 4**, which was assumed to be constant(Zhu & Yu, 2015), or ignored(Shi & Yin, 2019) in previous work.

Results from **Chapter 3** and **4** show that type I and II error rates are consistently matched for all analyses, regardless of the Bayesian priors used. Frequentist and Bayesian methods differ philosophically, however, this dissertation shows that these two methods can be unified methodologically in group sequential trials, which can serve as a cornerstone of statistical analysis in clinical trials. The unified approach also lowers the entry for clinical trialists to use Bayesian methods in clinical trials, because the approach is developed using relationships commonly seen in classical clinical trials. A further investigation in **Chapter 5** demonstrates that Bayesian prior specifications, as well as prior sample size, allocation ratio and the number of analyses planned in

clinical trials, can affect the final trial samples sizes. Therefore, frequentist and Bayesian methods can outperform each other under specific circumstances. It is suggested for clinical trialists to use the proposed unified framework at the trial planning stage to help make appropriate decisions on choosing Bayesian or frequentist approaches.

6.2 Future Work

There are several limitations to this dissertation. First, only binary endpoints are considered in this dissertation, while other regular endpoints, such as normal outcomes and time-to-event, are not discussed. Second, the methods currently are restricted to clinical trials with two samples, although proposed unified approaches have the potential to be extended to multi-arm multi-stage clinical trials. Last, the normal approximation to the transformed posterior probability might not be accurate if the posterior probability is extremely high (e.g. 0.9999999), as the distribution of posterior probability will be truncated at the upper boundary of one.

Thus, we aim to tackle these limitations in the future through corresponding improvements. First, proposed unified approaches will be extended to clinical trials with normal and time-to-event endpoints. Whether the same type I and II error rates can be achieved will also be evaluated. Second, a multiplicity adjustment procedure will be incorporated into the unified approaches, to account for multiple comparisons in multi-arm clinical trials. Last, a truncated normal approximation will be developed to approximate the truncated posterior probability distribution.

REFERENCE

Aitkin, M. (1997). The Calibration of P-values, Posterior Bayes Factors and the AIC from the Posterior Distribution of the Likelihood. *Statistics and Computing*, *7*(4), 253–261. https://doi.org/10.1023/A:1018550505678

Altham, P. M. E. (1969). Exact Bayesian Analysis of a 2 × 2 Contingency Table, and Fisher's "Exact" Significance Test. *Journal of the Royal Statistical Society. Series B (Methodological)*, *31*(2), 261–269. Retrieved from http://www.jstor.org/stable/2984209

Anderson, K. M., & Clark, J. B. (2010). Fitting Spending Functions. *Statistics in Medicine*, *29*(3), 321–327. https://doi.org/10.1002/sim.3737

Anscombe, F. J., & Aumann, R. J. (1963). A Definition of Subjective Probability. *Ann. Math. Statist.*, *34*(1), 199–205. https://doi.org/10.1214/aoms/1177704255

Armitage, P, McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A*, *132*. https://doi.org/10.2307/2343787

Armitage, Peter. (1958). Sequential Methods in Clinical Trials. *American Journal of Public Health and the Nations Health*, *48*(10), 1395–1402. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1551774/

Bayarri, M. J., & Berger, J. O. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statist. Sci.*, *19*(1), 58–80. https://doi.org/10.1214/088342304000000116

Berger, J. O., & Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values. *Journal of the American Statistical Association*, *82*(397), 112–122.

Berry, D. A. (2006). Bayesian Clinical Trials. *Nature Reviews Drug Discovery*, *5*, 27. Retrieved from http://dx.doi.org/10.1038/nrd1927

Berry, D. A., & Stangl, D. P. (1996). Bayesian Methods in Health-related Research. In *Bayesian Biostatistics*.

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian Adaptive Methods for Clinical Trials* (First Edition). CRC Press.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)*, *143*(4), 383–430.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-sided Testing Problem. *Journal of American Statistical Association*, *7*(1), 207–216. https://doi.org/10.1007/BF02565110

Chen, L. M., Ibrahim, J. G., & Chu, H. (2014). Flexible Stopping Boundaries When Changing Primary Endpoints after Unblinded Interim Analyses. *Journal of Biopharmaceutical Statistics*, *24*(4), 817–833. https://doi.org/10.1080/10543406.2014.901341

Cohen, H. W. (2011). P Values: Use and Misuse in Medical Literature. *American Journal of Hypertension*, *24*(1), 18–23. Retrieved from http://dx.doi.org/10.1038/ajh.2010.205

Cornfield, J. (1966). Sequential Trials, Sequential Analysis and the Likelihood Principle. *The American Statistician*, *20*(2), 18–23. https://doi.org/10.1080/00031305.1966.10479786

Cornfield, J. (1969). The Bayesian Outlook and Its Application. *Biometrics*, *25*(4), 617–657.

Cornfield, J. (1976). Recent Methodological Contributions to Clinical Trials. *American Journal of Epidemiology*, *104*(4), 408–421. https://doi.org/10.1093/oxfordjournals.aje.a112313

Cornfield, J., & Greenhouse, S. W. (1967). On Certain Aspects of Sequential Clinical Trials. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Biology and Problems of Health* (pp. 813–829). Berkeley, Calif.: University of California Press. Retrieved from https://projecteuclid.org/euclid.bsmsp/1200513830

DasGupta, A., & Vidakovic, B. (1997). Sample Size Problems in ANOVA Bayesian Point of View. *Journal of Statistical Planning and Inference*, *65*(2), 335–347. https://doi.org/https://doi.org/10.1016/S0378-3758(97)00056-6

De Santis, F. (2007). Using Historical Data for Bayesian Sample Size Determination. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *170*(1), 95–113. https://doi.org/10.1111/j.1467-985X.2006.00438.x

DeGroot, M. H. (1973). Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio. *Journal of the American Statistical Association*, *68*(344), 966–969. https://doi.org/10.1080/01621459.1973.10481456

DeMets, D. L., & Lan, K. K. G. (1994). Interim Anlaysis: The Alpha Spending Function Approach. *Statistics in Medicine*, *13*(13–14), 1341–1352. https://doi.org/10.1002/sim.4780131308

Dempster, A. P. (1973). The Direct Use of Likelihood for Significance Testing. In *In Proceedings of Conference on Foundational Questions in Statistical Inference* (pp. 335–354). Aaarhus, Denmark: Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

Dersimonian, R. (1996). Meta Analysis in the Design and Monitoring of Clinical Trials. *Statistics in Medicine*, *15*(12), 1237–1248. https://doi.org/10.1002/(SICI)1097-0258(19960630)15:12<1237::AID-SIM301>3.0.CO;2-N

Diamond, G. A., & Kaul, S. (2004). Prior Convictions: Bayesian Approaches to the Analysis and Interpretation of Clinical Megatrials. *Journal of the American College of Cardiology*, *43*(11), 1929–1939. https://doi.org/10.1016/j.jacc.2004.01.035

Dickey, J. M. (1977). Is the Tail Area Useful as an Approximate Bayes Factor? *Journal of the American Statistical Association*, *72*(357), 138–142. https://doi.org/10.1080/01621459.1977.10479922

Efron, B. E. (2005). Bayesians, Frequentists, and Scientists. *Journal of the American Statistical Association*, *100*(469), 1–5. https://doi.org/10.1198/016214505000000033

Fayers, P. M., Ashby, R. D., & Parmar, M. K. B. (1997a). Tutorial in Biostatistics: Bayesian Data Monitoring in Clinical Trials. *Statistics in Medicine*, *16*, 1413–1430.

FDA. (2014). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

Feigelson, E. D., Loredo, T. J., & Building, S. S. (1992). The Promise of Bayesian Inference for Astrophysics, *297*, 275–297.

Fox, B. L. (1966). A Bayesian Approach to Reliability Assessment. *Santa Monica, CA: RAND Corporation*.

Freedman, L. S., & Spiegelhalter, D. J. (1989). Comparison of Bayesian with Group Sequential Methods for Monitoring Clinical Trials. *Controlled Clinical Trials*, *10*(1989), 357–367. https://doi.org/10.1016/0197-2456(89)90001-9

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy* . https://doi.org/10.3390/e19100555

Gönen, M. (2009). Bayesian Clinical Trials: No More Excuses. *Clinical Trials*, *6*(3), 203–204. https://doi.org/10.1177/1740774509105374

Goodman, S. (2008). A Dirty Dozen : Twelve P-Value Misconceptions. *Seminars in Hematology*, *3*(45), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greenland, S. (2006). Bayesian Perspectives for Epidemiological Research: I. Foundations and Basic Methods. *International Journal of Epidemiology*, *35*(3), 765–775. Retrieved from http://dx.doi.org/10.1093/ije/dyi312

Greenland, S., & Poole, C. (2013). Living with P Values: Resurrecting a Bayesian Perspective on Frequentist Statistics. *Epidemiology*, *24*(1). Retrieved from https://journals.lww.com/epidem/Fulltext/2013/01000/Living_with_P_Values__Resurrecting_a_Bayesian.9.aspx

Grieve, A. P., & Wiley, J. (2016). Idle Thoughts of a ‘ Well-calibrated ’ Bayesian in Clinical Drug Development, (January). https://doi.org/10.1002/pst.1736

Gross, A. J. (1971). The Application of Exponential Smoothing to Reliability Assessment. *Technometrics*, *13*(4), 877–883. https://doi.org/10.1080/00401706.1971.10488859

Gsponer, T., Gerber, F., Bornkamp, B., Ohlssen, D., Vandemeulebroecke, M., & Schmidli, H. (2014). A Practical Guide to Bayesian Group Sequential Designs. *Pharmaceutical Statistics*, *13*(1), 71–80. https://doi.org/doi:10.1002/pst.1593

Gupta, S. K. (2012). Use of Bayesian Statistics in Drug Development: Advantages and Challenges. *International Journal of Applied & Basic Medical Research*, *2*(1), 3–6. https://doi.org/10.4103/2229-516X.96789

Hartley, H. O., & Fitch, E. R. (1951). A Chart for the Incomplete Beta-function and the Cumulative Binomial Distribution. *Biometrika*, *38*(3–4), 423–426.

https://doi.org/10.1093/biomet/38.3-4.423

Haybittle, J. L. (1971). Repeated Assessment of Results in Clinical Trials of Cancer Treatment. *The British Journal of Radiology*, *44*(526), 793–797. https://doi.org/10.1259/0007-1285-44-526-793

Horra, J. D. La. (2005). Reconciling Classical and Prior Predictive P-Values in the Two-Sided Location Parameter Testing Problem. *Communications in Statistics - Theory and Methods*, *34*(3), 575–583. https://doi.org/10.1081/STA-200052129

Hughes, M. D. (1993). Reporting of Bayesian Analyses of Clinical Trials. *Statistics in Medicine*, *12*, 1651–1663. https://doi.org/10.1177/009286159102500308

Ibrahim, J. G., & Chen, M. (2000). Power Prior Distributions for Regression Models. *Statistical Science*, *15*(1), 46–60.

ICH. (2017). E9 (R1) Estimands and Sensitivity Analysis in Clinical Trials. *Guidance*, *9*(June).

Inoue, L. Y. T., Berry, D. A., & Parmigiani, G. (2005). Relationship Between Bayesian and Frequentist Sample Size Determination. *The American Statistician*, *59*(1), 79–87. https://doi.org/10.1198/000313005X21069

Irony, T. (n.d.). *The Value of Bayesian Approaches in the Regulatory Setting: Lessons from the Past and Perspectives for the Future*.

Jeffreys, H. (1961). *Theory of Probability*. *Theory of Probability* (Third). Oxford, England: Oxford.

Jennison, C., & Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. *Boca Raton: Chapman and Hall*. https://doi.org/10.1016/s0197-2456(00)00127-6

Joseph, L., du Berger, R., & Bélisle, P. (1997). Bayesian and Mixed Bayesian/Likelihood Criteria for Sample Size Determination. *Statistics in Medicine*, *16*(7), 769–781. https://doi.org/10.1002/(SICI)1097-0258(19970415)16:7<769::AID-SIM495>3.0.CO;2-V

Kass, R. E. (2011). Statistical Inference: The Big Picture. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *26*(1), 1–9. https://doi.org/10.1214/10-STS337

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Katsis, A., & Toman, B. (1999). Bayesian Sample Size Calculations for Binomial Experiments. *Journal of Statistical Planning and Inference*, *81*(2), 349–362. https://doi.org/https://doi.org/10.1016/S0378-3758(99)00019-1

Kawasaki, Y., Shimokawa, A., Yamada, H., & Miyaoka, E. (2016). A Bayesian Equivalency Test for Two Independent Binomial Proportions. *Journal of Biopharmaceutical Statistics*, *26*(4), 781–789. https://doi.org/10.1080/10543406.2015.1074919

Lachin, J. M. (1981). Introduction to Sample Size Determination and Power Analysis for Clinical Trials. *Controlled Clinical Trials*, *2*(2), 93–113. https://doi.org/https://doi.org/10.1016/0197-2456(81)90001-5

Lai, T. L., Lavori, W., P., & Tsang, K. W. (2015). Adaptive Design of Confirmatory Trials: Advances and Challenges. *Contemporary Clinical Trials*, *33*(4), 395–401. https://doi.org/10.1038/nbt.3121.ChIP-nexus

Lan, G. K. ., & Demets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials. *Biometrika*, *70*(3), 659–663. https://doi.org/10.1093/biomet/70.3.659

Laptook, A. R., Shankaran, S., Tyson, J. E., Munoz, B., Bell, E. F., Goldberg, R. N., … Higgins, R. D. (2017). Effect of Therapeutic Hypothermia Initiated After 6 Hours of Age on Death or Disability among Newborns with Hypoxic-ischemic Encephalopathy a Randomized Clinical Trial. *JAMA - Journal of the American Medical Association*, *318*(16), 1550–1560. https://doi.org/10.1001/jama.2017.14972

Lecoutre, B. (2011). 11 - The Bayesian Approach to Experimental Data Analysis. In C. R. Rao, J. P. Miller, & D. C. B. T.-E. S. M. for M. S. Rao (Eds.), *Essential Statistical Methods for Medical Statistics* (pp. 308–344). Boston: North-Holland. https://doi.org/https://doi.org/10.1016/B978-0-444-53737-9.50014-1

Lee, J. J., & Chu, C. T. (2012). Bayesian Clinical Trials in Action. *Statistics in Medicine*. https://doi.org/10.1002/sim.5404

Lee, M. D. (2006). A Hierarchical Bayesian Model of Human Decision-Making on an Optimal Stopping Problem. *Cognitive Science*, *30*(3), 1–26. https://doi.org/10.1207/s15516709cog0000_69

Lehmacher, W., & Wassmer, G. (1999). Adaptive Sample Size Calculations in Group Sequential Trials Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics*, *55*(4), 1286–1290.

Lewis, R. J., Lipsky, A. M., & Berry, D. A. (2007). Bayesian Decision-theoretic Group Sequential Clinical Trial Design based on a Quadratic Loss Function: a Frequentist Evaluation. *Clinical Trials*, *4*(1), 5–14. https://doi.org/10.1177/1740774506075764

Lindley, D. V. (1957). A Statisitcal Paradox. *Biometrika*, *44*(1–2), 187–192. https://doi.org/10.1093/biomet/44.1-2.187

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511662973

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist Roadmap. *American Statistician*, *60*(3), 213–223. https://doi.org/10.1198/000313006X117837

M'Lan, C. E., Joseph, L., & Wolfson, D. B. (2008). Bayesian Sample Size Determination for Binomial Proportions. *Bayesian Analysis*, *3*(2), 269–296. https://doi.org/10.1214/08-BA310

Marden, J. I. (2000). Hypothesis Testing: From p Values to Bayes Factors. *Journal of the American Statistical Association*, *95*(452), 1316–1320. https://doi.org/10.2307/2669779

Micheas, A. C., & Dey, D. K. (2003). Prior and Posterior Predictive P-Values in the One-Sided Location Parameter Testing Problem. *Sankhyā: The Indian Journal of Statistics (2003-2007)*, *65*(1), 158–178. Retrieved from http://www.jstor.org/stable/25053252

Micheas, A. C., & Dey, D. K. (2007). Reconciling Bayesian and Frequentist Evidence in the One-Sided Scale Parameter Testing Problem. *Communications in Statistics - Theory and Methods*, *36*(6), 1123–1138. https://doi.org/10.1080/03610920601076610

Moatti, M., Zohar, S., Facon, T., Moreau, P., Mary, J.-Y., & Chevret, S. (2013). Modeling of Experts' Divergent Prior Beliefs for a Sequential Phase III Clinical Trial. *Clinical Trials*, *10*(4), 505–514. https://doi.org/10.1177/1740774513493528

Morita, S., Thall, P. F., & Peter, M. (2012). Prior Effective Sample Size in Conditionally Independent Hierarchical Models. *Bayesian Analysis*, *7*(3), 591–614. https://doi.org/10.1214/12-BA720

Müller, P., Berry, D. A., Grieve, A. P., & Krams, M. (2006). A Bayesian Decision-Theoretic Dose-Finding Trial. *Decision Analysis*, *3*(4), 197–207. https://doi.org/10.1287/deca.1060.0079

Novick, M. R., & Grizzle, J. E. (1965). A Bayesian Approach to the Analysis of Data from Clinical Trials. *Journal of the American Statistical Association*, *60*(309), 81–96. https://doi.org/10.2307/2283139

O'Brien, P. C., & Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*, *35*(3), 549–556. https://doi.org/10.2307/2530245

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in Clinical Trial Design. *Pharmaceutical Statistics*, *4*(3), 187–201. https://doi.org/10.1002/pst.175

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., … Smith, P. G. (1976). Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient. I. Introduction and design. *British Journal of Cancer*, *34*(6), 585–612. https://doi.org/10.1038/bjc.1976.220

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, *64*(2), 191–199. Retrieved from http://dx.doi.org/10.1093/biomet/64.2.191

Pratt, J. W. (1965). Bayesian Interpretation of Standard Inference Statements. *Journal of the Royal Statistical Society. Series B (Methodological)*, *27*(2), 169–203.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Harvard University Graduate School of Business Administration (Division of Research); Bailey & Swinfen.

Rosner, G. L., & Berry, D. A. (1995). A Bayesian Group Sequential Design for a Multiple Arm Randomized Clinical Trial. *Statistics in Medicine*, *14*(4), 381–394. https://doi.org/doi:10.1002/sim.4780140405

Rubin, D. B. (1983). Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

Sahu, S. K., & Smith, T. M. F. (2006). A Bayesian Method of Sample Size Determination with Practical Applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(2), 235–253. https://doi.org/10.1111/j.1467-985X.2006.00408.x

Samaniego, F. J., & Reneau, D. M. (1994). Toward a Reconciliation of the Bayesian and

Frequentist Approaches to Point Estimation. *Journal of the American Statistical Association*, *89*(427), 947–957. https://doi.org/10.1080/01621459.1994.10476828

Saville, B. R., Connor, J. T., Ayers, G. D., & Alvarez, J. (2014). Monitoring of Clinical Trials. *Clinical Trials*, *11*(4), 485–493. https://doi.org/10.1177/1740774514531352.The

Schüler, S., Kieser, M., & Rauch, G. (2017). Choice of futility boundaries for group sequential designs with two endpoints. *BMC Medical Research Methodology*, *17*(1), 119. https://doi.org/10.1186/s12874-017-0387-4

Sellke, T., Bayarri, M. J., & Berger, J. O. (2007). Calibration of p Values for Testing Precise Null Hypotheses, *55*(1), 62–71.

Shi, H., & Yin, G. (2019). Control of Type I Error Rates in Bayesian, (2), 399–425.

Shuirmann, D. J. (1987). A Comparison of the Two Oone-sided Tests Procedure and the Power. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.

Smith, I., & Ferrari, A. (2014). Equivalence between the Posterior Distribution of the Likelihood Ratio and a P-value in an Invariant Frame. *Bayesian Analysis*, *9*(4), 939–962. https://doi.org/10.1214/14-BA877

Spiegelhalter, David J. (2004). Incorporating Bayesian Ideas into Health-Care Evaluation. *Statistical Science*, *19*(1), 156–174. https://doi.org/10.1214/088342304000000080

Spiegelhalter, David J, Freedman, L. S., & Parmar, M. K. B. (1994). Bayesian Approaches to Randomized Trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 357–416. https://doi.org/10.2307/2983527

Teira, D. (2011). Frequentist versus Bayesian Clinical Trials. In F. B. T.-P. of M. Gifford (Ed.), *Handbook of the Philosophy of Science* (Vol. 16, pp. 255–297). Amsterdam: North-Holland. https://doi.org/https://doi.org/10.1016/B978-0-444-51787-6.50010-6

Thall, P. F., & Estey, E. H. (1993). A Bayesian Strategy for Screening Cancer Treatments Prior to Phase II Clinical Evaluation, *12*(April 1992), 1197–1211.

Thall, P. F., Simon, R. M., & Estey, E. H. (1995). Bayesian Sequential Monitoring Designs for Single-arm Clinical Trials with Multiple Outcomes. *Statistics in Medicine*, *14*(4), 357–379. https://doi.org/10.1002/sim.4780140404

Ventz, S., & Trippa, L. (2014). Bayesian Designs and the Control of Frequentist Characteristics: A Practical Solution. *Biometrics*, *71*(1), 218–226. https://doi.org/10.1111/biom.12226

Wang, H., Chow, S. C., & Chen, M. (2005). A Bayesian Approach on Sample Size Calculation for Comparing Means. *Journal of Biopharmaceutical Statistics*, *15*(5), 799–807. https://doi.org/10.1081/BIP-200067789

Weiss, R. (1997). Bayesian Sample Size Calculations for Hypothesis Testing. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *46*(2), 185–191. Retrieved from http://www.jstor.org/stable/2988523

Whitehead, J., & Stratton, I. (1983). Group Sequential Clinical Trials with Triangular Continuation Regions. *Biometrics*, *39*(1), 227–236.

Whitehead, J., Valdés-Márquez, E., Johnson, P., & Graham, G. (2008). Bayesian Sample Size for Exploratory Clinical Trials Incorporating Historical Data. *Statistics in Medicine*, *27*(13), 2307–2327. https://doi.org/doi:10.1002/sim.3140

Wu, Y., Shih, W. J., & Moore, D. F. (2008). Elicitation of a Beta Prior for Bayesian Inference in Clinical Trials, *50*, 212–223. https://doi.org/10.1002/bimj.200710390

Xie, F., Ji, Y., & Tremmel, L. (2012). A Bayesian Adaptive Design for Multi-dose, Randomized, Placebo-controlled Phase I/II Trials. *Contemporary Clinical Trials*, *33*(4), 739–748. https://doi.org/https://doi.org/10.1016/j.cct.2012.03.001

Yin, Y. (2011). Generalized P-values and Bayesian Evidence in the One-sided Testing Problems under Exponential Distributions. *Statistica Neerlandica*, *65*(3), 319–336. https://doi.org/10.1111/j.1467-9574.2011.00487.x

Yin, Y., & Wang, B. (2016). The Agreement between the Generalized p Value and Bayesian Evidence in the One-Sided Testing Problem. *International Journal of Mathematics and Mathematical Sciences*, *2016*, 1–7. https://doi.org/10.1155/2016/8656909

Yin, Y., & Zhao, J. (2013). Testing Normal Means: the Reconcilability of the P value and the Bayesian Evidence. *The Scientific World Journal*, *2013*, 381539. https://doi.org/10.1155/2013/381539

Yu, Z., Ramakrishnan, V., & Meinzer, C. (2019). Simulation optimization for Bayesian multi-arm multi-stage clinical trial with binary endpoints. *Journal of Biopharmaceutical Statistics*, *29*(2), 306–317. https://doi.org/10.1080/10543406.2019.1577682

Yusuf, S., & Flather, M. (1995). Magnesium in Acute Myocardial Infarction. *BMJ*, *310*(6982), 751 LP – 752. https://doi.org/10.1136/bmj.310.6982.751

Zaslavsky, B. G. (2010). Bayesian Versus Frequentist Hypotheses Testing in Clinical Trials with Dichotomous and Countable Outcomes. *Journal of Biopharmaceutical Statistics*, *20*(5), 985–997. https://doi.org/10.1080/10543401003619023

Zaslavsky, B. G. (2012). Bayesian Hypothesis Testing in Two-Arm Trials with Dichotomous Outcomes. *Biometrics*, *69*(1), 157–163. https://doi.org/10.1111/j.1541-0420.2012.01806.x

Zaslavsky, B. G., & Scott, J. (2012). Sample Size Estimation in Single-Arm Clinical Trials with Multiple Testing Under Frequentist and Bayesian Approaches. *Journal of Biopharmaceutical Statistics*, *22*(4), 819–835. https://doi.org/10.1080/10543406.2012.676585

Zhu, H., & Yu, Q. (2015). A Bayesian Sequential Design Using Alpha Spending Function to Control Type I Error. *Statistical Methods in Medical Research*, *26*(5), 2184–2196. https://doi.org/10.1177/0962280215595058